



Research article

Probabilistic versus deterministic machine learning for credit default prediction: A multidimensional benchmarking study under severe class imbalance

André Engan¹, Lars Ole Hjelkrem² and Morten Risstad^{1,*}

¹ Norwegian University of Science and Technology (NTNU), Dept. of Industrial Economics and Technology Management, Trondheim, Norway

² Sparebanken Møre, Ålesund, Norway

* **Correspondence:** Email: morten.risstad@ntnu.no; Tel: +47-97166263.

Abstract: Deploying machine learning models for credit decisions under regulatory mandates for explainability and reliability requires evaluation frameworks beyond predictive accuracy. Yet, benchmarking studies in credit risk still focus narrowly on the area under the receiver operating characteristic curve (AUC), neglecting probability calibration, uncertainty decomposition, and interpretability. We address this gap by benchmarking ten models, spanning deterministic and probabilistic families, across four evaluation dimensions on a proprietary dataset of 50,819 customers from a Norwegian financial institution. Each customer is represented by 124 consecutive daily account balance observations and a binary 12-month default flag. Across 180 configurations combining five feature engineering modules with four resampling strategies, variational inference-based Bayesian models (Bayesian neural network, Bayesian logistic regression, convolutional-Bayesian neural network) exhibit posterior collapse on this imbalanced tabular dataset, producing near-uniform predicted probabilities and near-zero epistemic uncertainty, particularly under oversampling. Natural gradient boosting (NGB) emerges as the most effective probabilistic model, matching random forest and XGBoost in test AUC (0.80) while providing informative aleatoric uncertainty estimates and economically interpretable SHAP attributions. We further show that feature engineering and resampling effects are non-additive and model-specific, indicating both as first-order design variables in credit risk pipelines. These findings extend the literature to probabilistic evaluation dimensions and provide practitioners with guidance for model selection in regulatory-compliant credit decision systems under similar data conditions. Note that our performance figures represent exploratory upper-bound benchmarks rather than strictly unbiased out-of-sample estimates.

Keywords: probabilistic machine learning; credit default prediction; uncertainty quantification; class imbalance; SHAP interpretability

1. Introduction

The deployment of machine learning models in high-stakes credit decisions is increasingly shaped by regulatory requirements that go beyond predictive performance. The European Union AI Act and

Basel III/IV supervisory guidelines mandate that models used in credit assessments must be explainable, reliable, and auditable, yet the machine learning credit risk literature almost exclusively evaluates models on AUC, ignoring probability calibration and uncertainty quantification [1, 2]. This disconnect between regulatory expectation and evaluation practice is particularly acute for probabilistic models, which are theoretically motivated by their capacity to quantify predictive uncertainty but are rarely evaluated on whether their uncertainty estimates are actually informative or trustworthy under the class imbalance conditions prevalent in credit data.

Machine learning has demonstrated strong empirical performance across a wide range of supervised learning tasks, including medical diagnosis [3, 4], asset pricing [5], time series analysis [6], volatility predictions [7] and credit risk assessment [8, 9]. In the credit domain, specifically, the increasing regulatory emphasis on model explainability and uncertainty quantification, reinforced by frameworks such as the European Union's AI Act and the SR 11-7 model risk guidance, has motivated evaluation frameworks that extend beyond predictive accuracy alone.

The motivating gap in the existing literature is precise: Despite substantial theoretical interest in probabilistic machine learning for uncertainty-aware prediction [10, 11], no systematic benchmarking study has evaluated both deterministic and probabilistic models on imbalanced financial tabular data across performance, calibration, uncertainty decomposition, and interpretability simultaneously. Benchmarking studies such as [8] and [12] provide valuable comparisons of deterministic models, but concentrate on predictive performance, overlooking probabilistic approaches. [13] documents failures of Bayesian deep learning under distributional shift, but the credit domain, characterized by severe class imbalance, tabular structure, and regulatory interpretability requirements, has not been systematically examined. This leaves practitioners without evidence-based guidance for whether probabilistic models actually deliver on their theoretical promise in this context [9].

In this paper, we benchmark ten machine learning models, logistic regression (LR), random forest (RF), eXtreme gradient boosting (XGB), deep neural network (DNN), convolutional-deep neural network (CNN-DNN), naive Bayes (NB), Bayesian logistic regression (BLR), Bayesian neural network (BNN), natural gradient boosting (NGB), and convolutional-Bayesian neural network (CNN-BNN), on a proprietary Norwegian bank dataset of 50,819 customers. Each customer is represented by 124 consecutive daily account balance observations and a binary default indicator; 819 customers (1.6%) defaulted within 12 months of the observation window. We evaluate all models across four dimensions, predictive performance, probability calibration, uncertainty decomposition, and SHAP-based interpretability, across 180 experimental configurations combining feature engineering modules and resampling strategies.

The results yield three primary findings. First, variational inference-based Bayesian models exhibit posterior collapse on this dataset, producing near-uniform predicted probabilities with near-zero epistemic uncertainty regardless of true class membership — a failure mode that is model-specific and robust to feature engineering choice. Second, NGB achieves competitive test AUC (0.80), provides informative aleatoric uncertainty estimates with broad distributional support, and yields economically interpretable SHAP attributions concentrated in intuitively plausible balance features. Third, feature engineering and resampling effects on model performance are non-additive and model-specific, with oversampling methods substantially degrading Bayesian model calibration in configurations that exhibit posterior collapse.

This paper makes four contributions:

- We provide the first systematic benchmarking of probabilistic machine learning models for credit default prediction across four evaluation dimensions: predictive performance, probability calibration, uncertainty decomposition, and SHAP-based interpretability.
- We document posterior collapse in variational inference-based Bayesian models on imbalanced tabular financial data, extending findings from [13] and [14] to the credit risk domain. We observe that posterior collapse is more severe under oversampling, consistent with oversampling being a contributing factor, though other architectural and initialization factors cannot be ruled out.
- We demonstrate that NGB provides reliable aleatoric uncertainty estimates competitive with state-of-the-art deterministic models, establishing it as a practical alternative for regulatory-compliant credit scoring.
- We show that feature engineering and resampling strategy effects are non-additive and model-specific, suggesting these may require co-optimization with model architecture in credit risk machine learning pipelines.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the data and experimental setup. Section 4 presents the methods. Section 5 reports results. Section 6 discusses findings and limitations. Section 7 concludes the paper.

2. Related work

2.1. Machine learning benchmarking in credit risk: A performance-focused literature

The machine learning credit risk literature has generated extensive evidence that ensemble methods outperform simpler models on standard classification metrics. Reference [8] benchmarks 41 classifiers and finds that random forest and gradient boosting machines consistently outperform logistic regression, long the standard in credit scoring [15, 16], while [12] and [17] replicate RF's dominance in corporate loan and consumer credit settings. References [18] and [19] demonstrate that CNN architectures can automatically extract temporal patterns from account balance sequences, eliminating manual feature engineering, while achieving competitive AUC values on Norwegian banking data directly comparable in structure to our dataset. These studies collectively establish a rich empirical baseline against which our deterministic model results can be situated.

However, this literature shares a common limitation: evaluation is almost exclusively restricted to AUC, accuracy, and threshold-dependent metrics such as F1-score and recall [8, 20, 21]. Probability calibration, whether predicted probabilities correspond to empirical default frequencies, is rarely assessed. Uncertainty decomposition into aleatoric and epistemic components is effectively absent from the credit risk benchmarking literature. Interpretability analyses using model-agnostic methods such as SHAP are increasingly common [22–24], but are not systematically paired with uncertainty assessments. The result is a literature that can rank models by discriminative ability but provides limited guidance for practitioners who must choose models for deployment in systems where calibrated probability estimates are required by contract or regulation.

2.2. Bayesian deep learning for uncertainty quantification: Promise and empirical failures

Bayesian neural networks and related probabilistic deep learning approaches are motivated by their theoretical capacity to distinguish between aleatoric uncertainty (irreducible noise in the data) and

epistemic uncertainty (reducible model uncertainty), providing more informative outputs than point-estimate classifiers [10, 25]. Monte Carlo dropout [11] and Bayes by backprop [26] have extended this capacity to practical deep learning architectures. Reference [27] reports that BNNs provide both predictive accuracy and feature relevance rankings in financial applications, while [28] demonstrate improved uncertainty estimates from convolutional-Bayesian hybrid models.

Nonetheless, empirical evidence increasingly documents systematic failures of these approaches. Reference [13] found that BNNs exhibit poor calibration under distributional shift, with variational inference-based approaches particularly susceptible to overconfident posterior approximations. Reference [14] surveyed the uncertainty quantification literature and identify posterior collapse, where the variational posterior converges to the prior, failing to encode data-dependent information, as a pervasive failure mode under aggressive resampling and class imbalance. Reference [29] documented that Bayesian deep ensembles often outperform variational BNNs in practice, suggesting that the variational approximation is frequently the binding constraint. These empirical findings motivate our investigation of whether posterior collapse manifests in credit default prediction under the severe class imbalance characteristic of this domain (1.6% default rate), and whether NGB, which does not rely on variational inference, provides a more reliable alternative.

2.3. *Multidimensional evaluation frameworks: An unmet need in credit risk ML*

A small but growing literature argues that credit risk machine learning evaluation must extend beyond AUC to encompass calibration and interpretability [30, 31]. Reference [32] found that aleatoric uncertainty dominates epistemic uncertainty in credit risk settings, suggesting that models which can represent aleatoric risk, may be more appropriate than those designed primarily for epistemic uncertainty quantification. Reference [33] established that many classifiers are systematically miscalibrated and require post-hoc correction, a finding that [34] extended to modern neural networks. Reference [35] provided the theoretical foundation for SHAP-based interpretability analyses, which [36] showed can be destabilized by class imbalance.

To our knowledge, no published benchmarking study has applied all four evaluation dimensions — performance, calibration, uncertainty decomposition, and interpretability, simultaneously to a credit default prediction task, evaluated probabilistic models including BNN, BLR, NGB, and CNN-BNN, and systematically varied both feature engineering and resampling as design variables. Our study is also the first, as far as we know, to document the specific interaction between ADASYN/SMOTE oversampling and variational BNN posterior collapse in a credit risk context, extending the general findings of [14] with model-specific and configuration-level evidence.

3. Data and experimental setup

3.1. *Dataset*

The data are a proprietary dataset provided by a Norwegian bank. They comprise 50,819 observations, each corresponding to an individual customer who has taken out a loan. The dataset consists of 126 variables: A unique customer identifier (excluded from modeling), 124 variables (*dg_1* to *dg_124*) representing daily account balances for the 124 days preceding the score date, and *M12* indicating whether the customer defaulted within 12 months after the score date. The bank anonymized the dataset via standardization,

which preserves the relational structure of the data without affecting model training or evaluation [37].

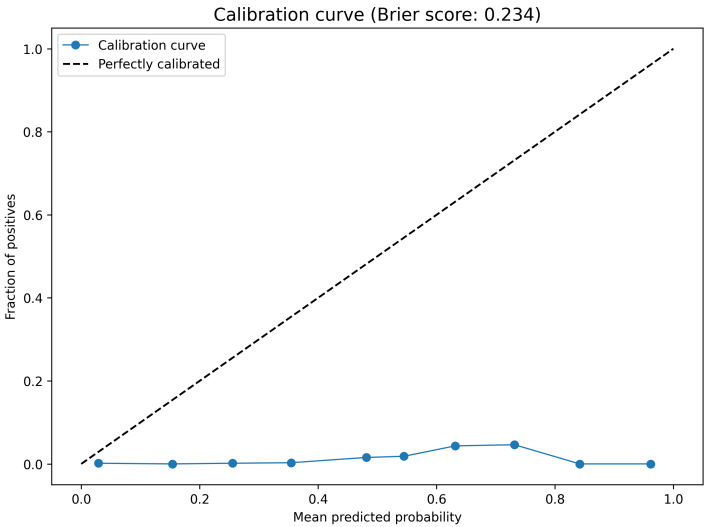


Figure 1. Class distribution. The dataset contains 50,000 non-defaulters and 819 defaulters (1.6% default rate).

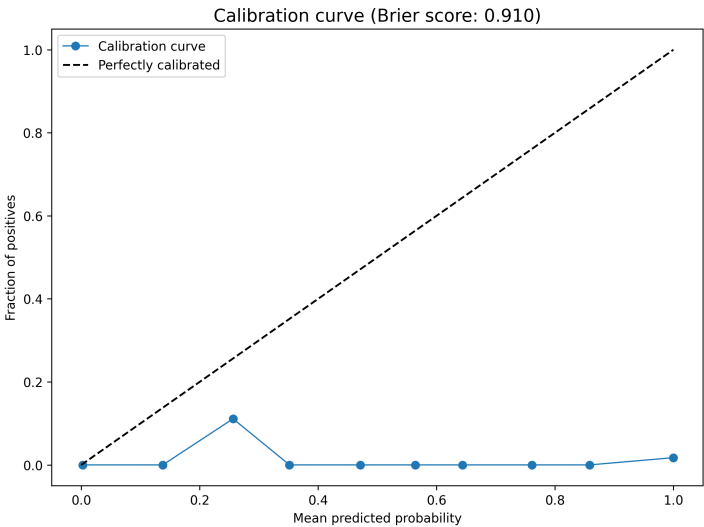


Figure 2. Average daily account balance by default status across the 124-day observation window. Defaulters maintain consistently lower balances; non-defaulters exhibit broader variation.

Figure 1 shows the class distribution: 50,000 non-defaulters and 819 defaulters, yielding a default rate of 1.6%. This severe imbalance is characteristic of retail credit data and directly motivates our systematic investigation of resampling strategies. Figure 2 illustrates that defaulters maintain consistently lower account balances than non-defaulters throughout the 124-day window, with non-defaulters exhibiting approximately twice the rolling standard deviation, providing a visual basis for the predictive signal in the data.

There are no missing values. Descriptive statistics for a subset of the raw features are reported in Table A1 in Appendix 7.

3.2. Feature engineering

To extract informative representations from the 124 raw daily balance features, we implement five feature engineering modules that can be combined flexibly within the preprocessing pipeline. A complete description of all engineered features is provided in Appendix 7.

Account statistics. Aggregated summary statistics computed across all 124 daily balances per customer: Total balance, minimum, maximum, mean, and standard deviation. These capture overall level and dispersion.

Rolling aggregates. Seven-day rolling statistics: mean, median, standard deviation, interquartile range, trend slope, and trend change count. These capture short-term dynamics that may signal financial distress.

Variability metrics. Features quantifying irregular behavior: sign changes in daily differences, proportion of days with positive change, spike count, maximum spike magnitude, and average rolling standard deviation.

SelectKBest. Dimensionality reduction retaining the top $k = 10$ features ranked by ANOVA F-statistic, selecting features with strong linear discriminative power.

SelectFromModel. Model-based selection using a random forest classifier's Gini importance scores, retaining features above the median importance threshold to capture non-linear feature–target relationships.

3.3. Resampling techniques

To address class imbalance, four resampling strategies are evaluated: Synthetic Minority Over-sampling Technique (SMOTE), ADaptive SYNthetic sampling (ADASYN), random oversampling (RO), and random undersampling (RU). These methods are applied exclusively to the training set to prevent data leakage [38, 39]. SMOTE and ADASYN expand the training set from 38,114 to approximately 75,000 observations; RO similarly duplicates minority instances; and RU contracts the training set to 1228 balanced observations by removing majority instances.

3.4. Experimental configurations

Preliminary experimentation identified three feature engineering configurations (Base, FE_3, and FE_5) that demonstrated the strongest per-model performance, as shown in Table 1. Resampling techniques were applied only to these three configurations, yielding 18 total configurations per model (Table 2). This design enables targeted assessment of how different preprocessing strategies interact with model architecture.

Table 1. Preliminary AUC (area under the ROC curve) by model and feature engineering module, prior to resampling. AUC values are computed on cross-validation held-out folds (training partition only, $n = 38,114$); the test set was not used. Base, FE_3, and FE_5 were selected as the configurations showing the strongest performance across the majority of models; FE_1, FE_2, and FE_4 are included for reference but were not carried forward to the resampling evaluation.

	LR	RF	XGB	DNN	CNN-DNN	NB	BNN	BLR	NGB	CNN-BNN
Base	0.58	0.70	0.72	0.70	0.70	0.54	0.50	0.58	0.71	0.50
FE_1	0.66	0.70	0.72	0.68	0.71	0.54	0.51	0.61	0.71	0.50
FE_2	0.63	0.70	0.73	0.66	0.66	0.61	0.50	0.59	0.72	0.50
FE_3	0.66	0.70	0.73	0.68	0.68	0.63	0.50	0.63	0.72	0.50
FE_4	0.65	0.64	0.72	0.71	0.71	0.65	0.52	0.65	0.71	0.50
FE_5	0.60	0.70	0.72	0.71	0.71	0.55	0.50	0.60	0.72	0.50

Table 2. Experimental configurations combining feature engineering modules and resampling techniques. Resampling was applied only to Base, FE_3, and FE_5 configurations; FE_1, FE_2, and FE_4 are included for reference in Table 1 but were not carried forward to the resampling stage. Abbreviations: SMOTE = Synthetic Minority Over-sampling TEchnique; ADASYN = ADaptive SYNthetic sampling; RO = Random Oversampling; RU = Random Undersampling.

Configuration	Description
Base	No feature engineering or resampling. Raw balance features retained.
FE_1	Account statistics, rolling aggregates, and variability metrics. No resampling.
FE_2	Account statistics and variability metrics. No resampling.
FE_3	Account statistics and rolling aggregates. No resampling.
FE_4	SelectKBest ($k = 10$). No resampling.
FE_5	SelectFromModel (RandomForestClassifier). No resampling.
Base + SMOTE	Base with SMOTE resampling.
Base + ADASYN	Base with ADASYN resampling.
Base + RO	Base with random oversampling.
Base + RU	Base with random undersampling.
FE_3 + SMOTE	FE_3 with SMOTE resampling.
FE_3 + ADASYN	FE_3 with ADASYN resampling.
FE_3 + RO	FE_3 with random oversampling.
FE_3 + RU	FE_3 with random undersampling.
FE_5 + SMOTE	FE_5 with SMOTE resampling.
FE_5 + ADASYN	FE_5 with ADASYN resampling.
FE_5 + RO	FE_5 with random oversampling.
FE_5 + RU	FE_5 with random undersampling.

3.5. Data splitting

The dataset is partitioned on a customer-wise (row-wise) basis into training (75%, $n = 38,114$) and test (25%, $n = 12,705$) sets using stratified sampling to preserve the default rate in both subsets. Stratified k -fold cross-validation with $k = 5$ is applied to the training set for model validation and hyperparameter selection. Figure 3 illustrates this procedure.

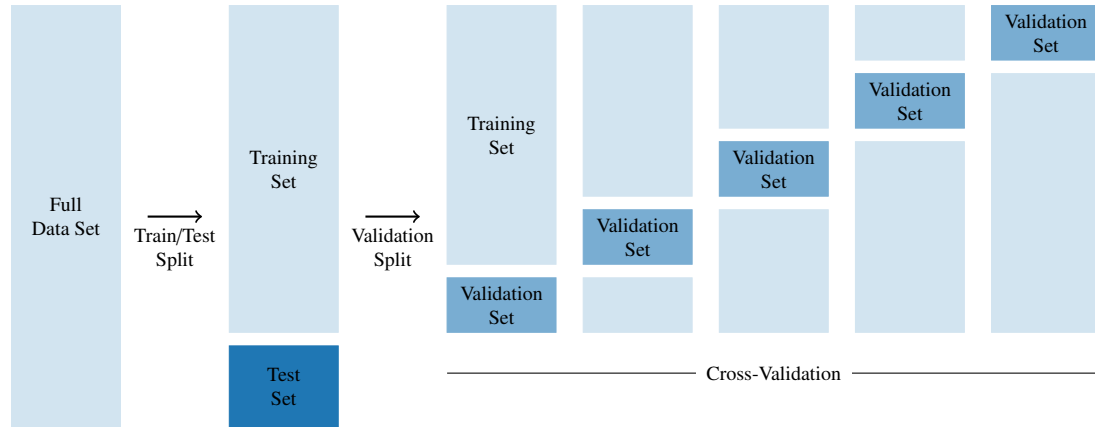


Figure 3. Data splitting procedure. The full dataset is first partitioned customer-wise into training (75%) and test (25%) sets. Stratified 5-fold cross-validation is applied to the training set for model validation.

3.6. Internal validity of the benchmarking design

Several aspects of the benchmarking design bear on internal validity. First, the partitioning is customer-wise (row-wise), ensuring that no customer appears in both the training and test sets; within-customer leakage is therefore not a concern given the cross-sectional structure of the observations. Second, hyperparameter tuning uses cross-validation AUC on the training set only. However, the best preprocessing configuration for each model is selected by test AUC across 18 configurations, which inflates reported performance relative to prospective deployment and constitutes a form of multiple testing on the test set; reported AUC values should therefore be treated as upper bounds on expected out-of-sample performance. Third, the comparison of BNN, BLR, and NGB is not a controlled experiment, the models differ in architecture, inference method, and training objective, and differences in performance and uncertainty estimates cannot be attributed to any single factor; the comparison is appropriately treated as a benchmarking study rather than a causal analysis. Fourth, uncertainty decomposition into aleatoric and epistemic components is defined differently across model families (NGB uses distributional parameters; BNN uses variational posteriors), and cross-model comparisons of uncertainty magnitudes are interpretive rather than strictly comparable.

3.7. Data limitations

The dataset carries several limitations that constrain generalisability. The proprietary and anonymized nature restricts external validation and limits the granularity of economic feature interpretation, as SHAP attributions reflect patterns in standardized rather than named features. The dataset contains only 124 daily account balance observations with no credit bureau variables, income, employment status,

or loan-to-value ratios; the high aleatoric uncertainty observed across all models may partly reflect genuine informational limits of this feature set rather than modeling failures. Single-institution data from one time period prevents out-of-time and out-of-sample validation. Finally, the 1.6% default rate requires careful metric interpretation: AUC and Brier score are our primary evaluation metrics given their robustness under imbalance, while precision, recall, and F1-score are reported for completeness, but are not the basis for model ranking.

4. Methods

4.1. Machine learning models

Ten models are evaluated: five deterministic (LR, RF, XGB, DNN, CNN-DNN) and five probabilistic (NB, BNN, BLR, NGB, CNN-BNN). The selection spans linear and non-linear, parametric and non-parametric, and shallow and deep architectures, enabling a systematic comparison across a wide range of modeling strategies.

Among the deterministic models, *logistic regression* (LR) estimates default probability via the logistic sigmoid applied to a linear combination of features, trained by maximizing the log-likelihood with optional L_2 regularization [40]. *Random forest* (RF) aggregates predictions from an ensemble of decision trees, each grown on a bootstrap sample with random feature subsets; the bagging mechanism reduces variance and promotes diversity without requiring an explicit global optimiser [41]. *eXtreme gradient boosting* (XGB) builds trees sequentially to correct residual errors, incorporating second-order gradients and a regularisation term on tree complexity [42]. The *deep neural network* (DNN) stacks multiple fully connected layers with dropout and batch normalization, trained via backpropagation and binary cross-entropy loss [43]. The *convolutional-DNN* (CNN-DNN) prepends 1D convolutional layers to the DNN, reshaping each input observation into a pseudo-time-series to extract local patterns before the dense classification head.

Among the probabilistic models, *naive Bayes* (NB) applies Bayes' theorem with the conditional independence assumption, yielding closed-form parameter estimates but potentially sacrificing discriminative power when features are correlated [40]. The *Bayesian neural network* (BNN) treats weights as random variables and approximates the posterior via mean-field variational inference (VI), producing a predictive distribution over outputs that decomposes into epistemic (weight) and aleatoric (data) uncertainty [44]. *Bayesian logistic regression* (BLR) applies the same VI framework to a linear classifier, assigning Gaussian priors to regression coefficients; the posterior is approximated by minimizing the Kullback-Leibler divergence $KL(q||p_{\text{post}})$ from the variational distribution q to the true posterior p_{post} . *Natural gradient boosting* (NGB) extends gradient boosting to predict full Bernoulli distributions, using natural gradients, scaled by the Fisher information matrix, to improve convergence stability and decompose aleatoric uncertainty from the predicted distribution parameters [45]. The *convolutional-BNN* (CNN-BNN) combines deterministic convolutional feature extraction with Bayesian dense layers, inheriting uncertainty quantification from the variational posterior while leveraging the CNN's capacity for local pattern detection.

4.2. Implementation and training

All models were implemented in Python 3.11 using a modular, reproducible codebase with random seed fixed to 42. LR, RF, and NB were implemented with scikit-learn; XGB with the XGBoost library; NGB with the NGBoost library; and DNN, CNN-DNN, BNN, BLR, and CNN-BNN with Keras and TensorFlow Probability. Resampling was performed with imbalanced-learn. Interpretability analysis used the SHAP library [35]. Unless otherwise stated, parameters follow library defaults.

All models share the same training set and are evaluated on the identical held-out test set. Neural networks were trained for 50–100 epochs depending on the configuration. BNN and BLR training maximizes the evidence lower bound (ELBO) (equivalently minimizes the negative ELBO) combining an expected log-likelihood term with a KL-regularisation term, $\text{KL}(q_\phi(\mathbf{w})||p(\mathbf{w}))$, that penalizes divergence of the variational posterior from the prior. NGB minimizes the negative log-likelihood of the Bernoulli predictive distribution using natural gradient steps.

4.3. Hyperparameter optimization

Hyperparameters were tuned via sequential model-based optimization (SMBO) using cross-validation AUC on the training set as the objective. The search ranges and selected values for all models are reported in Appendix 7. Briefly: RF used 450 estimators at depth 10; XGB used 250 estimators, learning rate 0.036, and depth 2; NGB used 400 estimators at learning rate 0.012. Neural architectures were evaluated across candidate layer sizes and dropout rates; the BNN's final architecture comprised two Bayesian dense layers of width 32.

4.4. Evaluation metrics

Model performance is assessed on the held-out test set using AUC and Brier score as primary metrics, supplemented by accuracy, precision, recall, and F1-score for diagnostic completeness. AUC is threshold-independent and robust under class imbalance; the Brier score, defined as the mean squared error between predicted probabilities and binary outcomes, jointly penalizes discrimination failure and probability miscalibration, making it particularly informative for probabilistic models [46]. Under the 1.6% default rate of this dataset, precision-recall AUC (PR-AUC; equivalently, average precision score) is more informative than ROC-AUC for minority-class evaluation, as it focuses the assessment on the positive class across all decision thresholds [47]. PR-AUC and threshold-optimized metrics (e.g., F1 at a prevalence-calibrated threshold, precision at top- k) were not computed in this study; we flag this as a limitation and recommend their inclusion in future benchmarking work. The Brier score, which is reported for all 180 configurations, provides a partial complement to AUC by jointly penalizing discriminative failure and probability miscalibration in a threshold-independent manner. Calibration curves (reliability diagrams) visualize the correspondence between predicted probabilities and observed event frequencies.

4.5. Uncertainty quantification

The terms ‘epistemic’ and ‘aleatoric’ uncertainty carry different operational meanings across model families in this study. Epistemic uncertainty in BNN and CNN-BNN is the variance of predicted probabilities across $T = 100$ MC forward passes through the variational posterior over weights; in BLR it is the variance across $T = 100$ MC samples from the posterior over coefficients. Aleatoric uncertainty

in BNN and CNN-BNN is the mean output variance $\frac{1}{T} \sum_{t=1}^T \hat{p}_t(1 - \hat{p}_t)$ across forward passes; in NGB it is the Bernoulli predictive variance $\hat{p}(1 - \hat{p})$, which is the minimal single-parameter characterization for binary outcomes. These quantities are derived from fundamentally different generative models and are not directly comparable in absolute magnitude across model families; cross-model comparisons should be treated as qualitative rather than quantitative [25].

Let $\hat{p}_t$ denote the model-predicted default probability on a given observation at Monte Carlo sample t , and let $\bar{p} = \frac{1}{T} \sum_{t=1}^T \hat{p}_t$ denote the mean predicted probability across T samples. For BNN and CNN-BNN, epistemic uncertainty is approximated by the variance of $\hat{p}_t$ across $T = 100$ Monte Carlo (MC) forward passes through the variational posteriors; aleatoric uncertainty is estimated as $\frac{1}{T} \sum_{t=1}^T \hat{p}_t(1 - \hat{p}_t)$ across the same $T = 100$ MC forward passes [25]. For NGB, aleatoric uncertainty is captured by the variance of the Bernoulli predictive distribution, which equals $\hat{p}(1 - \hat{p})$ where $\hat{p}$ is the predicted default probability; NGB does not natively decompose epistemic uncertainty. For BLR, epistemic uncertainty is estimated as the variance across $T = 100$ MC samples from the posterior over coefficients. Summary statistics (mean, median, standard deviation, interquartile range [IQR]) and kernel density estimates (KDE, using Gaussian kernel with Scott's rule bandwidth) are reported for each model.

4.6. Interpretability

SHAP values are computed for all models using a random sample of 500 test observations, providing model-agnostic feature attributions grounded in cooperative game theory [35]. TreeSHAP is applied for tree-based models (RF, XGB, NGB); KernelSHAP for the remaining models. Global feature importance is summarized via mean absolute SHAP values (bar plots); distributional attribution patterns are visualized via beeswarm plots.

5. Results

5.1. Predictive performance

Table 3 presents test set performance for the best-performing configuration of each model, selected by test AUC. Figures 4–6 provide complementary diagnostic plots. Selected results are provided in Appendix C1; the full 180-row table is available from the authors on request.

Confusion matrices and derived metrics (precision, recall, F1-score) are computed at a fixed threshold of 0.5. Under the prevailing 1.6% default rate, this threshold corresponds to severe over-prediction of the positive class for any probabilistically calibrated model, which explains the near-zero test precision values observed across all models. Operational deployment would require threshold calibration based on asymmetric decision costs or prevalence adjustment; such analysis is beyond the scope of this benchmarking study. AUC and Brier score, being threshold-independent, remain the primary comparative metrics.

Table 3. Test set performance ($n = 12,705$; 205 defaulters) for the best-performing configuration of each model, selected by test AUC. Precision, recall, and F1-score are computed at a classification threshold of 0.5. Brier score is reported for the test set only. Bold values indicate the best score per column across all models. Configuration codes refer to Table 2. Full results for all configurations appear in Table D1.

	Accuracy		Precision		Recall		F1		AUC		Brier
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Test
<i>Deterministic</i>											
LR (FE_3+SMOTE)	0.64	0.60	0.61	0.02	0.80	0.61	0.68	0.05	0.70	0.69	0.23
RF (FE_1)	0.81	0.78	0.08	0.05	0.96	0.64	0.15	0.08	0.94	0.80	0.11
XGB (FE_1)	0.63	0.64	0.04	0.04	0.87	0.81	0.07	0.07	0.84	0.80	0.18
DNN (FE_3+RO)	0.65	0.62	0.67	0.03	0.61	0.78	0.60	0.06	0.74	0.76	0.18
CNN-DNN (FE_3+RO)	0.66	0.59	0.65	0.03	0.71	0.83	0.67	0.06	0.73	0.76	0.19
<i>Probabilistic</i>											
NB (FE_4)	0.10	0.09	0.02	0.02	0.99	1.00	0.03	0.03	0.66	0.68	0.91
BNN (FE_5+ADASYN)	0.53	0.12	0.47	0.02	0.80	1.00	0.55	0.04	0.70	0.72	0.26
BLR (FE_4)	0.47	0.27	0.02	0.02	0.65	0.95	0.04	0.04	0.66	0.65	0.25
NGB (FE_3+RO)	0.80	0.68	0.76	0.04	0.89	0.76	0.82	0.07	0.89	0.80	0.17
CNN-BNN (FE_3+ADASYN)	0.65	0.31	0.60	0.02	0.92	0.96	0.72	0.04	0.74	0.74	0.26

RF emerges as the best-performing model overall, attaining the highest test AUC (0.80), accuracy (0.78), precision (0.05), F1-score (0.08), and lowest Brier score (0.11), a finding consistent with the broad credit scoring literature [8, 48]. Consistent with the theoretical expectation that bagging reduces variance, RF shows robustness to class imbalance in this dataset while preserving sensitivity to non-linear feature interactions. XGB matches RF in test AUC (0.80) but achieves higher recall (0.81 vs. 0.64) at the cost of substantially worse Brier score (0.18) and lower precision, suggesting a trade-off between recall and calibration, though the mechanism was not directly examined [42]. LR achieves an AUC of 0.69, consistent with its limited capacity to capture non-linear interactions in high-dimensional data; the high recall (0.61) but near-zero precision (0.02) reflect the model's tendency toward over-prediction under SMOTE oversampling. DNN and CNN-DNN attain moderate AUCs (0.76 for both), with high test recall but universally low precision; the convolutional layer adds no discriminative gain, consistent with findings by [49] that standard neural networks tend to underperform on tabular data without structural inductive bias.

Among the probabilistic models, NGB stands out as the only one achieving competitive performance, matching RF and XGB at test AUC 0.80 while delivering a Brier score of 0.17, superior to all other probabilistic models. BNN, BLR, and CNN-BNN exhibit a characteristic pattern: near-perfect recall (≥ 0.95), near-zero precision (≤ 0.02), and Brier scores of ≥ 0.25 . The AUC values of BNN (0.72) and BLR (0.65) indicate weak discriminative ability despite high recall, suggesting predictions are concentrated near the 0.5 threshold rather than differentiating high- from low-risk customers. NB performs poorly on all metrics except recall (1.00), with a Brier score of 0.91 reflecting completely unreliable probability estimates; the conditional independence assumption is strongly violated by the correlated daily balance features.

The calibration curves in Figure 6 reveal systematic overconfidence across all models: calibration curves lie below the diagonal for nearly the full range of predicted probabilities, indicating that predicted probabilities overestimate the true event rate. RF is partially calibrated at high predicted probabilities, which is operationally most important in default prediction where correctly identifying the highest-risk customers is the central task. NGB's calibration curve deviates more uniformly, without the partial recovery at high probabilities, signaling systematic overconfidence for high-risk customers. The Bayesian models, BNN, BLR, and CNN-BNN, fail to assign probabilities above 0.6 on the test set, further reflecting posterior collapse, discussed in Section 5.3.*

The universal recall-precision imbalance underscores the fundamental challenge posed by the 1.6% default rate. High AUC values coexist with near-zero F1-scores for most models, confirming that AUC alone is insufficient to characterize operational utility under class imbalance [47]. These results argue for a dual evaluation standard, AUC for discriminative capacity and Brier score for probabilistic reliability, applied conjointly.

*Confidence intervals on individual calibration bins are not shown. Given the test set of $n = 12,705$ observations with 205 defaulters (1.6%), bins at high predicted probabilities contain few observations and the displayed curves should be interpreted with caution in that region. Future work should include bin sample sizes and Wilson confidence intervals to improve diagnostic reliability.

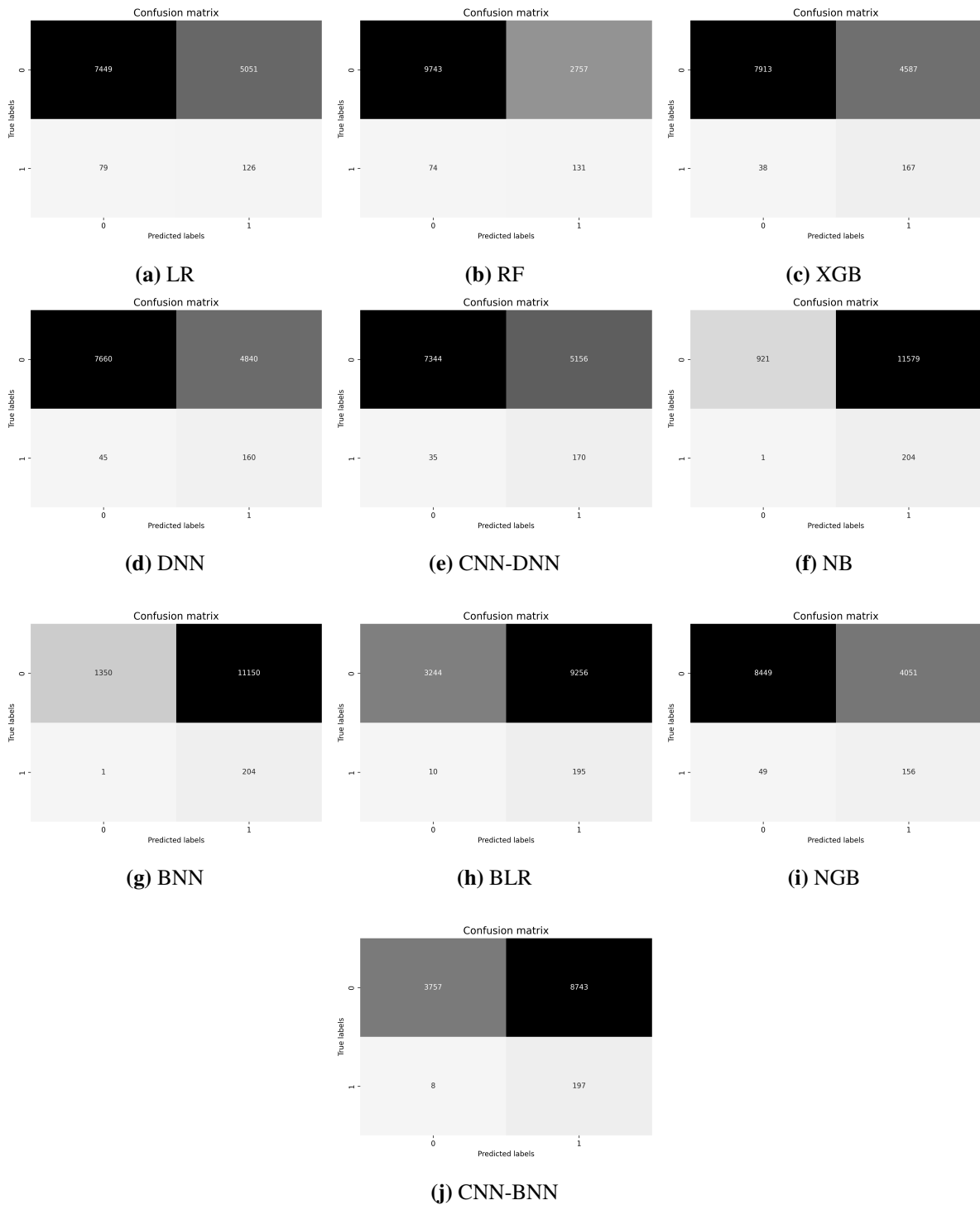


Figure 4. Confusion matrices on the held-out test set for the best-performing configuration of each model. Columns: (a) LR, (b) RF, (c) XGB, (d) DNN, (e) CNN-DNN, (f) NB, (g) BNN, (h) BLR, (i) NGB, (j) CNN-BNN.

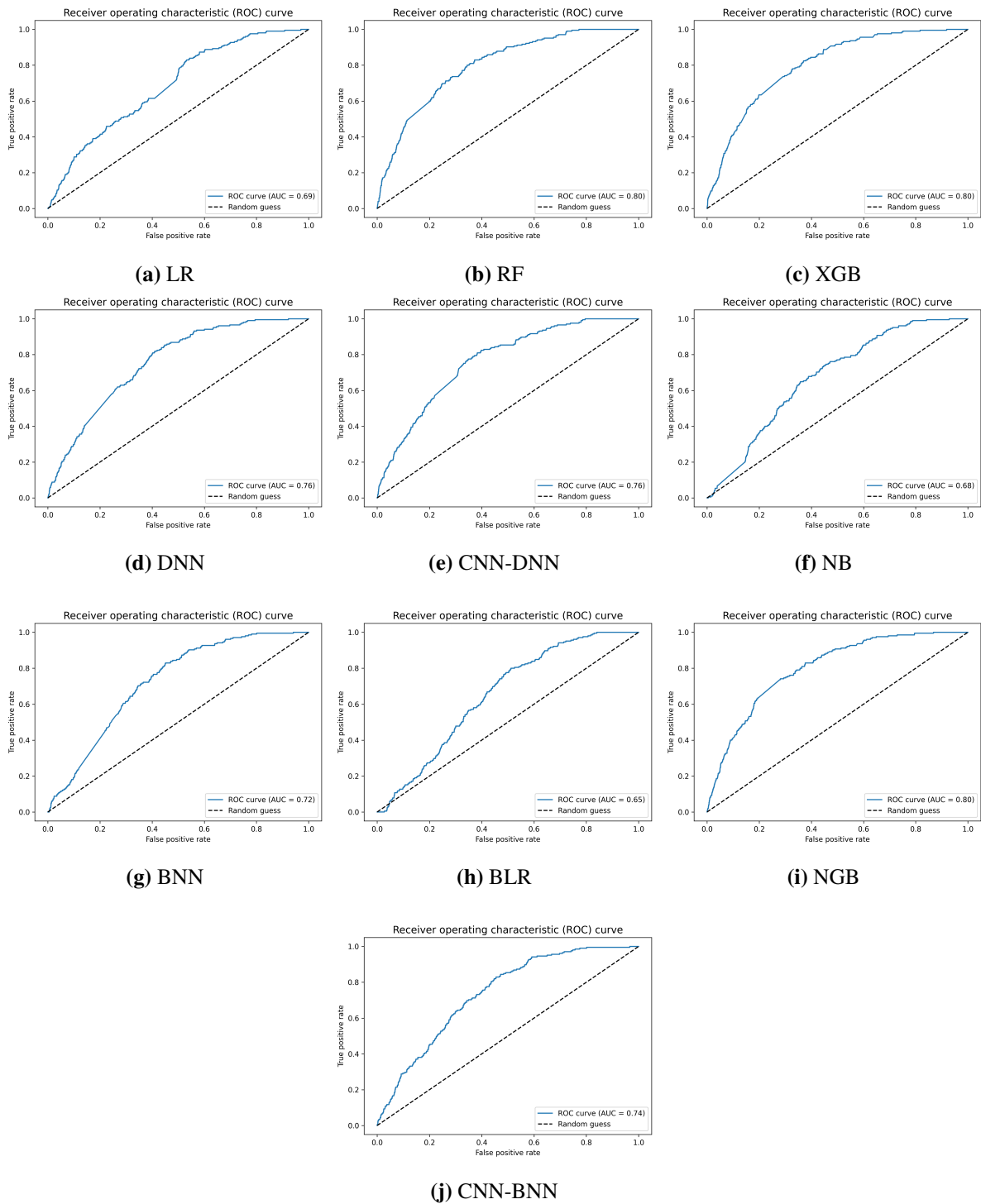


Figure 5. ROC curves on the held-out test set for the best-performing configuration of each model. Panels correspond to the same model ordering as Figure 4.

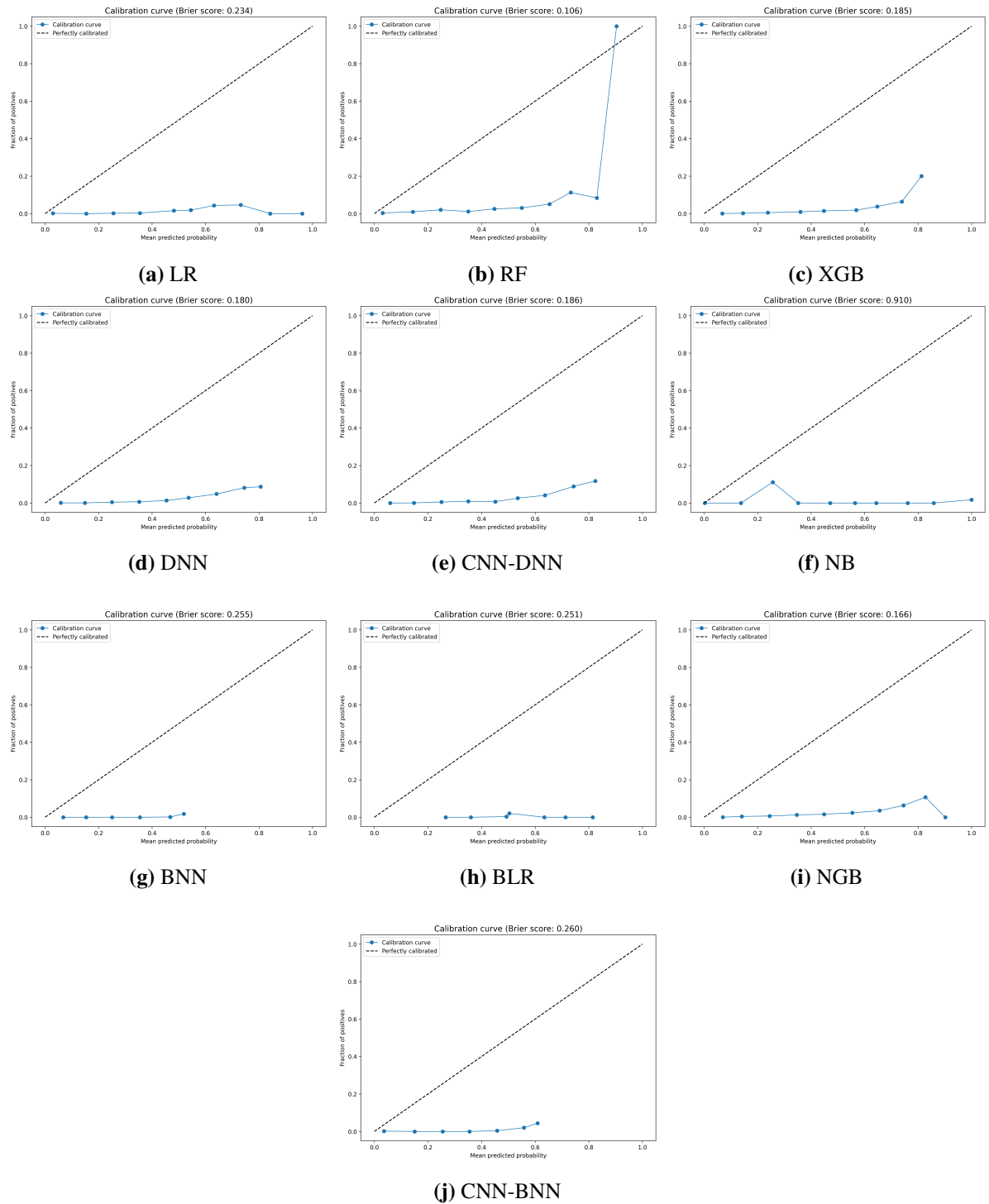


Figure 6. Calibration curves on the held-out test set for the best-performing configuration of each model. The diagonal indicates perfect calibration. Panels correspond to the same model ordering as Figure 4.

5.2. Effect of preprocessing on performance and robustness

Table 4 presents the variance in test AUC and Brier score across resampling methods for three representative feature engineering configurations per model. This detailed preprocessing analysis reveals that the choice of feature engineering and resampling strategy exerts a profound and model-specific influence on predictive reliability.

Table 4. Variance in test AUC and Brier score across resampling methods for each model across three feature engineering configurations (Base, FE_3, and FE_5). For each model, population variance (denominator $N = 5$) is computed across five resampling conditions: Baseline (no resampling), SMOTE, ADASYN, RO, and RU. Test set: $n = 12,705$ (205 defaulters). Individual configuration values appear in Table D1.

	Variance in AUC	Variance in Brier score
<i>LR</i>		
Base	0.000500	0.000470
FE_3	0.000050	0.000020
FE_5	0.000430	0.000120
<i>RF</i>		
Base	0.000220	0.001230
FE_3	0.000130	0.000870
FE_5	0.000170	0.001230
<i>XGB</i>		
Base	0.000330	0.000030
FE_3	0.000030	0.000030
FE_5	0.000270	0.000030
<i>DNN</i>		
Base	0.000030	0.000730
FE_3	0.000370	0.000730
FE_5	0.000030	0.000430
<i>CNN-DNN</i>		
Base	0.000230	0.001480
FE_3	0.001400	0.000830
FE_5	0.000200	0.002450
<i>NB</i>		
Base	0.000030	0.000180
FE_3	0.000020	0.000120
FE_5	0.000030	0.000170
<i>BNN</i>		
Base	0.008880	0.108720
FE_3	0.011030	0.107430
FE_5	0.010350	0.104420
<i>BLR</i>		
Base	0.007830	0.000000
FE_3	0.000850	0.000050
FE_5	0.003150	0.000000
<i>NGB</i>		
Base	0.000350	0.000420
FE_3	0.000100	0.000230
FE_5	0.000170	0.000280
<i>CNN-BNN</i>		
Base	0.008870	0.101670
FE_3	0.010930	0.104720
FE_5	0.011120	0.100480

The ensemble models, RF, XGB, and NGB, exhibit uniformly low variance in both AUC and Brier score (all below 0.0015 and 0.0013, respectively), indicating robustness to resampling-induced

distribution shifts. In contrast, BNN and CNN-BNN display extreme instability: AUC variance ranges from 0.009 to approximately 0.011 and Brier score variance from 0.100 to 0.109, two to three orders of magnitude larger than the tree-based models. These values indicate that BNNs are highly sensitive to the synthetic minority class introduced by SMOTE and ADASYN; whether this sensitivity reflects properties of the variational inference process specifically or interacting factors (e.g., learning rate, prior scale) cannot be determined from this benchmarking study alone. BLR shows instability in AUC variance (up to 0.008), but near-zero Brier variance, reflecting its tendency to collapse to a fixed probability regardless of input, discussed further in Section 5.3.

Feature engineering configurations FE_1 and FE_3, which introduce rolling statistics and distributional aggregates over the 124-day balance sequence — consistently support superior performance in tree-based models, yielding the best AUC and Brier scores for RF, XGB, and NGB. Deep learning models exhibit more modest and volatile gains from feature engineering, consistent with findings by [49] that standard architectures lack the inductive bias needed to exploit engineered tabular features. Random oversampling (RO) provides the most stable resampling strategy across model types; SMOTE and ADASYN improve recall in underfitting-prone models but are associated with elevated posterior instability and miscalibration in variational Bayesian models, while random undersampling (RU) degrades calibration in data-hungry architectures. These findings suggest that, on this dataset, feature engineering is associated with more consistent performance improvements than resampling strategy, and that optimal preprocessing must be co-designed with the target architecture rather than applied as a modular step.

5.3. Uncertainty quantification

Table 5 reports summary statistics of epistemic and aleatoric uncertainty for the four probabilistic models that provide uncertainty decomposition. Figure 7 visualizes the corresponding KDE distributions.

Table 5. Summary statistics of predictive uncertainty for the best-performing configuration of each probabilistic model. Epistemic uncertainty (variance units) is estimated as the variance of predictions across $T = 100$ MC forward passes for BNN and CNN-BNN, and across $T = 100$ MC samples from the posterior for BLR. Aleatoric uncertainty (probability units, range $[0, 1]$) is estimated as $\frac{1}{T} \sum_{t=1}^T \hat{p}_t(1 - \hat{p}_t)$ for BNN and CNN-BNN, and as $\hat{p}(1 - \hat{p})$ for NGB. NB is excluded as it does not provide uncertainty decomposition. Dashes indicate that a given uncertainty type is not applicable for the model. Test set: $n = 12,705$ (205 defaulters).

	Mean	Median	Std	IQR
<i>BNN</i>				
Epistemic	0.02883	0.02483	0.01727	0.00150
Aleatoric	0.49691	0.49887	0.02055	0.00004
<i>BLR</i>				
Epistemic	0.00381	0.00163	0.00943	0.00022
Aleatoric	—	—	—	—
<i>NGB</i>				
Epistemic	—	—	—	—
Aleatoric	0.39964	0.42837	0.09336	0.16738
<i>CNN-BNN</i>				
Epistemic	0.03884	0.03674	0.01512	0.01748
Aleatoric	0.47054	0.49344	0.07953	0.00641

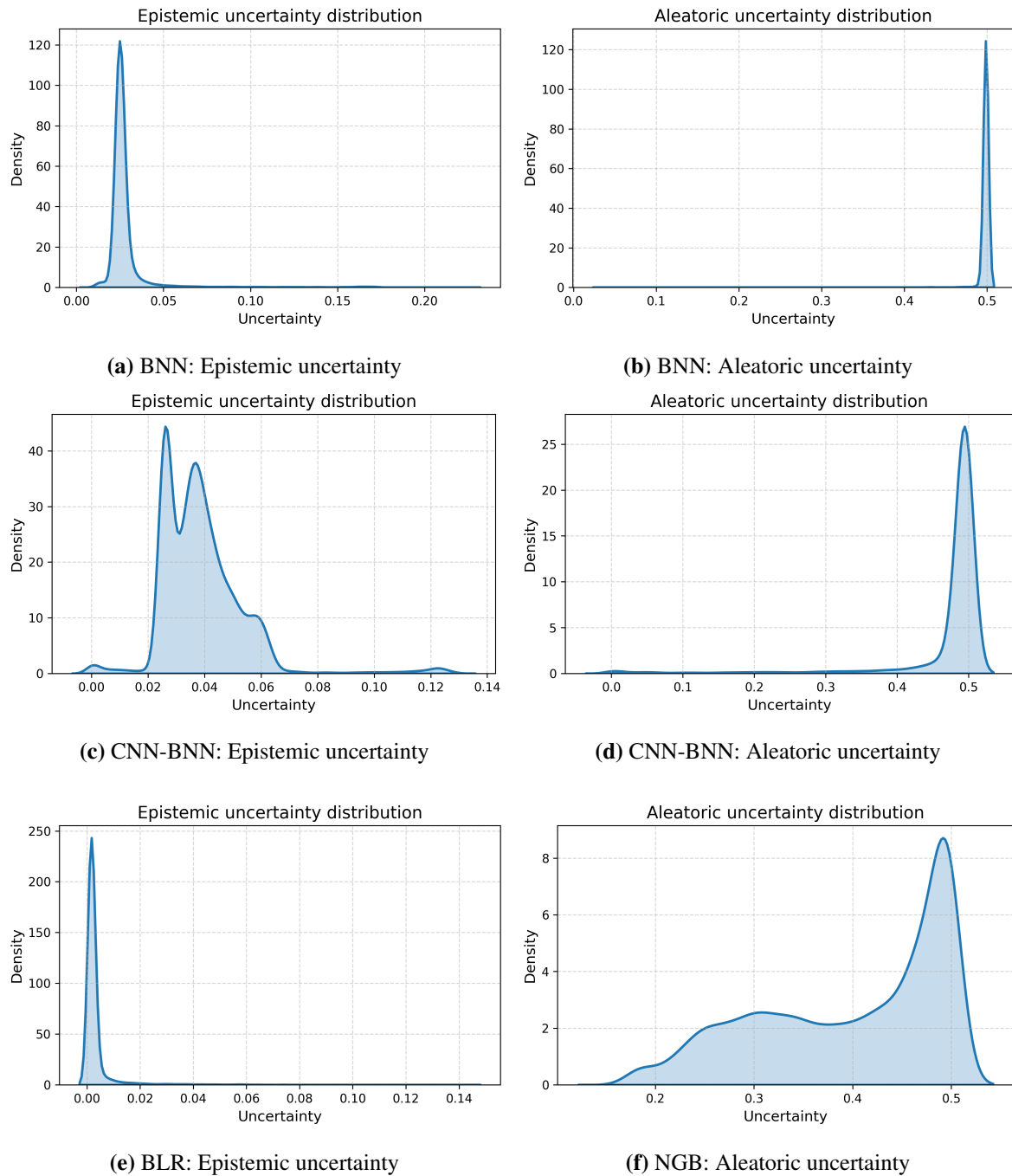


Figure 7. KDE plots of predictive uncertainty for the best-performing configuration of each probabilistic model. Panels (a and b): BNN epistemic and aleatoric uncertainty. Panels (c and d): CNN-BNN epistemic and aleatoric uncertainty. Panel (e): BLR epistemic uncertainty. Panel (f): NGB aleatoric uncertainty.

The BNN reveals a striking pattern: Low epistemic uncertainty (mean 0.029, IQR 0.0015) coexists with near-constant aleatoric uncertainty (mean 0.497, IQR 0.00004). The negligible IQR of aleatoric uncertainty indicates that the model assigns essentially the same predicted probability, approximately

0.5, to every observation in the test set, regardless of input. The synthetic minority-class samples introduced by ADASYN oversampling may exacerbate this failure: interpolated synthetic instances have low epistemic diversity and distort the variational objective by emphasising belief mass near 0.5, as documented by [14] and [29]. Reference [50] shows that BNNs can remain stable under moderate imbalance with appropriate variational inference strategies, further suggesting that the collapse observed here is specific to the combination of mean-field VI, default priors, and aggressive oversampling rather than an inherent property of Bayesian neural networks. This behavior, combined with the low epistemic spread, is a hallmark of what we term posterior collapse: the model produces near-uniform predictions consistent with the variational posterior failing to update meaningfully from the prior. We use this term descriptively to refer to the observed near-uniform prediction behavior; we cannot formally verify from output behavior alone that the variational posterior has converged to the prior [26, 51]. Table 6 quantifies this directly: BNN assigns a mean predicted probability of 0.501 for the positive class against an empirical rate of 0.016, yielding an absolute discrepancy of 0.485.

Table 6. Comparison of empirical class priors and mean predicted class probabilities for BNN, BLR, and CNN-BNN on the test set ($n = 12,705$; 205 defaulters). Mean predicted $P(y = 1)$ is taken over $T = 100$ MC forward passes per observation for BNN and CNN-BNN, and over $T = 100$ MC posterior samples for BLR. NGB is excluded as it does not exhibit the near-uniform prediction behavior documented here. An absolute difference near 0.5 indicates that the model predicts approximately uniformly regardless of the true class distribution.

Model	Empirical $P(y = 1)$	Mean predicted $P(y = 1)$	Abs. diff. ($y = 1$)	Empirical $P(y = 0)$	Mean predicted $P(y = 0)$	Abs. diff. ($y = 0$)
BNN	0.016	0.501	0.485	0.984	0.499	0.485
BLR	0.016	0.501	0.485	0.984	0.499	0.485
CNN-BNN	0.016	0.493	0.477	0.984	0.507	0.477

CNN-BNN shows slightly higher epistemic mean (0.039) and broader IQR (0.0175), suggesting that the convolutional front-end introduces more variance into the approximate posterior. Aleatoric uncertainty remains high (mean 0.471), but with more dispersion than BNN (Std 0.080 vs. 0.021), indicating some residual input-sensitivity despite the overall posterior collapse evident in Table 6. BLR displays the lowest epistemic uncertainty of all models (mean 0.004), reinforcing the earlier observation that it has converged to a near-degenerate posterior driven by the prior. Without aleatoric decomposition, BLR effectively behaves as a deterministic classifier despite its Bayesian formulation.

As noted in Section 4.5, NGB’s aleatoric uncertainty ($\hat{p}(1 - \hat{p})$) and the BNN/CNN-BNN uncertainty estimates are not directly comparable in absolute magnitude; the following cross-model discussion is therefore qualitative.

NGB is the only probabilistic model producing relatively informative and input-sensitive uncertainty estimates, in the sense that its aleatoric distribution spans a substantially wider range than those of the Bayesian models. Its aleatoric uncertainty (mean 0.300, IQR 0.167) spans a substantially wider range than the Bayesian models, reflecting heteroskedastic risk assessment across the customer population. The broad IQR indicates that NGB assigns meaningfully different uncertainty levels to different observations, a property essential for credible risk-sensitive decision-making, rather than collapsing to a uniform near-0.5 estimate. The KDE plot in Figure 7(f) confirms this multimodal uncertainty distribution, with distinct clusters corresponding to low- and high-uncertainty prediction regimes. These results are consistent with [45], who show that natural gradients decouple variance estimation from the mean prediction,

enabling more stable uncertainty modeling. Reference [32] found, in a related default-prediction setting, that aleatoric uncertainty dominates across all models, indicating that the dominant source of prediction uncertainty reflects genuine data variability rather than model limitations. NGB's exclusive focus on aleatoric uncertainty is therefore well-matched to this characteristic of default prediction data, where irreducible outcome ambiguity, arising from unobserved borrower characteristics and behavioral complexity, constitutes the primary source of predictive uncertainty [32, 52].

The alignment between calibration and uncertainty is inconsistent across models. In BNN, BLR, and CNN-BNN, the overconfident calibration curves (Figure 6) coincide with narrow uncertainty distributions, indicating that model certainty is not functioning as a corrective mechanism. NGB, however, demonstrates a more plausible risk–uncertainty alignment: wider aleatoric spread corresponds to the model's moderate but non-trivial miscalibration, enabling partial discrimination between high- and low-certainty predictions. These findings challenge the assumption that mean-field variational inference with default priors reliably produces calibrated and informative uncertainty estimates under severe class imbalance [10, 25]; in practice, poor inference quality can reduce probabilistic outputs to uninformative noise.

5.4. Interpretability

Figure 8 presents SHAP bar and beeswarm plots for the deterministic models; Figure 9 presents the same for the probabilistic models.

The tree-based models, RF, XGB, and NGB, exhibit stable, coherent, and economically interpretable feature attributions. All three assign the highest global importance to *min_balance*, *total_balance*, and *mean_balance*, with low values of these features consistently associated with higher predicted default probability. This directional pattern aligns with established credit risk intuition: customers with chronically low or declining account balances present elevated default risk [53, 54]. NGB additionally assigns high importance to *std_balance*, with wider SHAP distributions than RF, reflecting nuanced modeling of heteroskedasticity that corresponds to its broader aleatoric uncertainty estimates. LR similarly concentrates attributions on *min_balance* and *total_balance*, with a broad distribution at low values driving positive predictions, consistent with its linear structure. DNN and CNN-DNN also prioritize *total_balance* and *min_balance*, with moderate but stable contributions from secondary features; the CNN layer adds no distinctive attribution structure, reinforcing the absence of meaningful local patterns in the data.

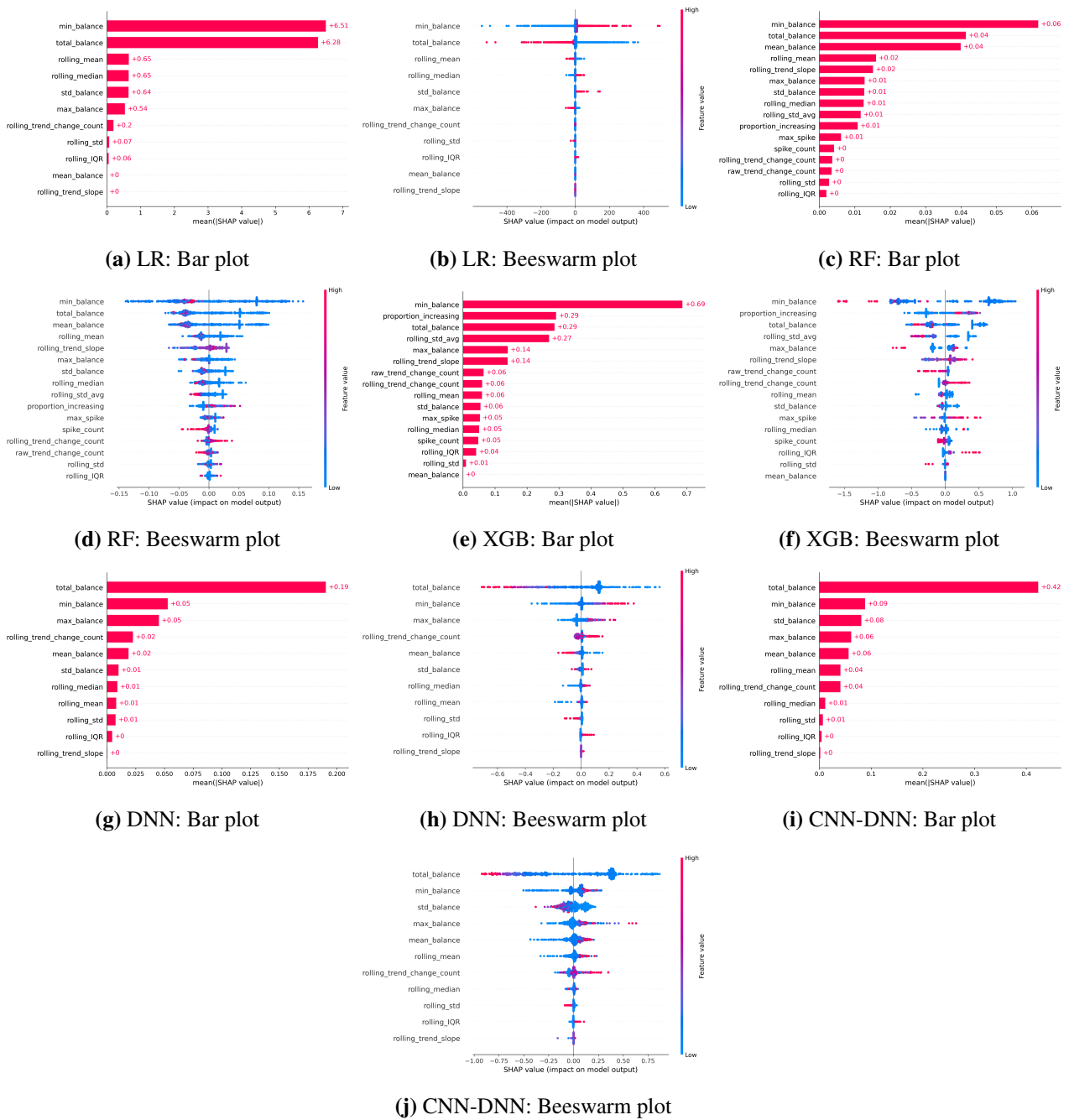


Figure 8. SHAP bar plots (mean absolute SHAP value) and beeswarm plots for the deterministic models. Bar plots indicate global feature importance; beeswarm plots show the distribution of individual SHAP values, with color indicating feature magnitude. SHAP values computed on 500 test instances.

BNN distributes attribution thinly and near-symmetrically across a handful of features (*dg_10*, *dg_4*, *dg_1*, *dg_6*), with uniformly low SHAP magnitudes indicating that the model has not learned stable feature-to-target associations, a direct consequence of the posterior collapse documented above. CNN-BNN concentrates attribution on *total_balance* with a wide but near-monotone SHAP distribution, suggesting a partially informative predictive signal despite overall posterior collapse. BLR highlights *dg_114* and *dg_109* with wide SHAP dispersion, likely reflecting high collinearity in the daily feature sequence and sensitivity to specific feature positions. NB produces near-zero SHAP values uniformly across all features, consistent with its failure to extract informative predictive signals under the violated independence assumption.

Chen et al. [36] specifically demonstrates that class imbalance significantly reduces the stability of SHAP-based model explanations: Attribution scores computed under extreme minority-class ratios tend to be noisier and less reproducible than those in balanced settings, a risk directly applicable to the 1.6% default rate of this dataset. Separately, we note several further limitations of the SHAP analysis in this study. KernelSHAP attribution stability under highly correlated inputs was not formally assessed; the 124 consecutive daily balance features are highly intercorrelated by construction, which may induce instability in KernelSHAP attributions for non-tree-based models. Attribution reproducibility across random seeds and SHAP sample sizes was not examined, and no comparison to alternative interpretability methods (e.g., LIME, integrated gradients) was conducted. SHAP attributions for BNN, BLR, and CNN-BNN should be interpreted with particular caution given the near-uniform predicted probabilities documented in Section 5.3. The attributions for these models do not reliably reflect learned feature-outcome associations and are presented for completeness rather than as operationally useful explanations.

Taken together, the interpretability analysis reinforces the ranking established by performance and uncertainty metrics: RF, XGB, and NGB provide both accurate predictions and economically coherent explanations, while the Bayesian models and NB either fail to learn meaningful attributions or produce attribution patterns that may not reliably reflect learned feature-outcome associations, given the near-uniform predicted probabilities documented above and the absence of formal SHAP stability validation.

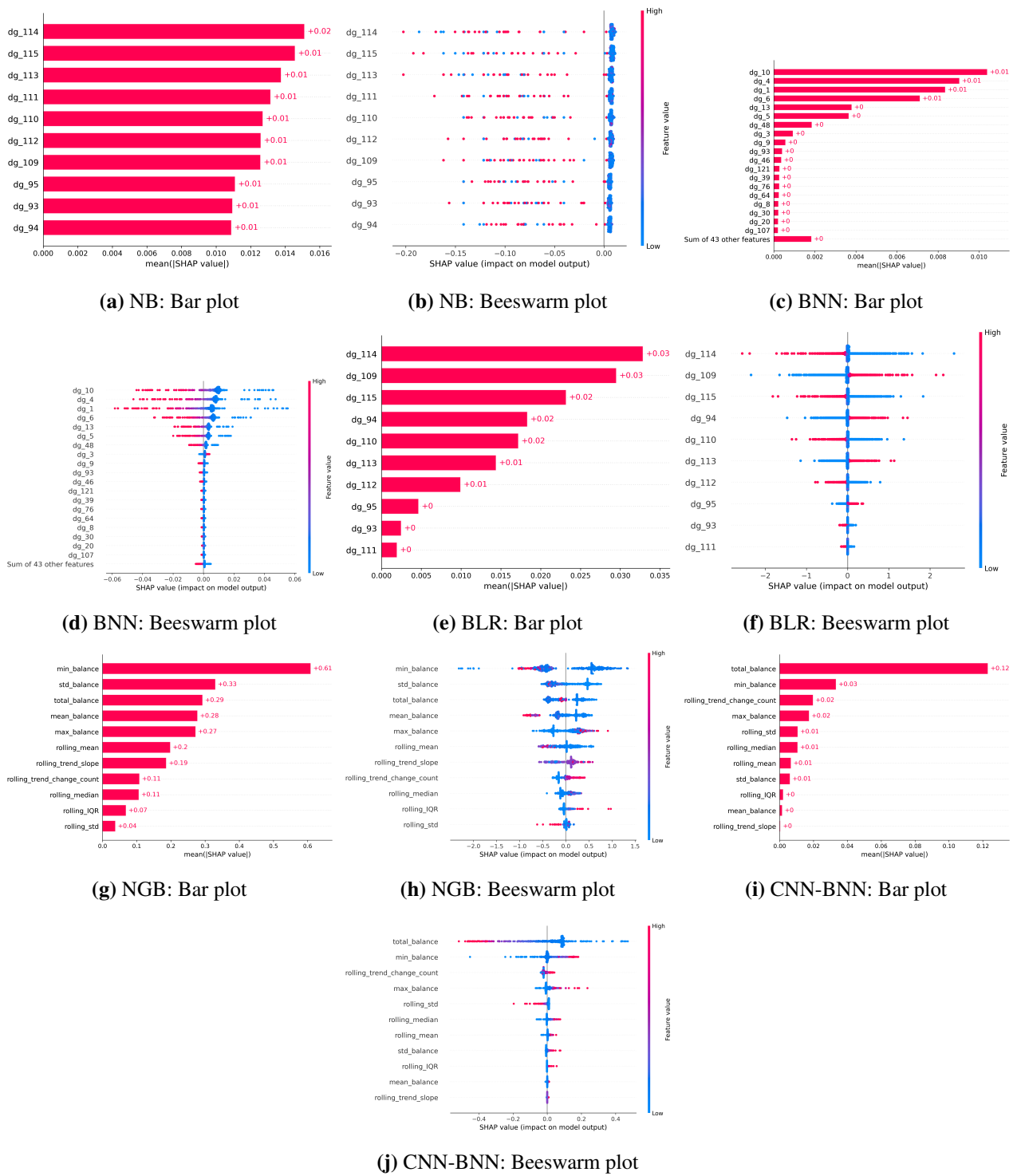


Figure 9. SHAP bar plots and beeswarm plots for the probabilistic models. Panels (a–b): NB, (c–d): BNN, (e–f): BLR, (g–h): NGB, (i–j): CNN-BNN. SHAP values computed on 500 test instances.

6. Discussion

We note that this is a benchmarking study and the observed associations between preprocessing choices, model architectures, and performance outcomes should not be interpreted causally. Controlled ablation experiments would be required to isolate the contribution of individual preprocessing steps or architectural features to the observed patterns.

6.1. NGB as the best-performing probabilistic model

Across all four evaluation dimensions, predictive performance, preprocessing robustness, uncertainty quantification, and interpretability, NGB consistently emerges as the most capable probabilistic model. Its test AUC (0.80) matches the best deterministic models (RF, XGB), while its Brier score (0.17) is substantially better than the remaining probabilistic models, though inferior to RF (0.11). The low variance in both AUC and Brier score across resampling conditions (Table 4) demonstrates that NGB's performance is structurally stable rather than configuration-dependent, a critical property for deployment in operational scoring systems where preprocessing pipelines may evolve. Its aleatoric uncertainty distribution (mean 0.300, IQR 0.167) is the only one among the probabilistic models to exhibit meaningful heterogeneity across predictions, providing credible risk differentiation rather than uniform near-0.5 estimates. Finally, its SHAP attributions are economically coherent: *min_balance*, *total_balance*, and *std_balance* drive predictions in directions consistent with established credit risk theory, and the attribution distributions are stable across test instances.

This constellation of properties reflects NGB's structural advantages for severely imbalanced tabular data. Unlike BNNs, which rely on mean-field VI and are vulnerable to posterior collapse when the minority class is sparse, NGB avoids explicit posterior inference altogether: its uncertainty emerges from the distributional parameters of a proper scoring-rule objective, insulating it from the prior-dominance failure mode observed here in BNN, BLR, and CNN-BNN. The natural gradient update simultaneously stabilizes convergence and improves variance modeling by incorporating information geometry into each boosting step, in contrast to the ordinary gradient steps of XGB, which prioritize discriminative accuracy over probabilistic calibration. Taken together, NGB represents a promising candidate for probabilistic credit scoring in settings similar to this study: it achieves discrimination on par with ensemble deterministic models while providing uncertainty estimates that are both informative and architecturally principled.

Nonetheless, NGB's limitations warrant acknowledgement. Its calibration curve lies below the diagonal uniformly, without the partial recovery observed for RF at high predicted probabilities; this implies systematic overconfidence for the highest-risk customers, precisely those for whom accurate probability estimates are most operationally important. Additionally, NGB does not decompose uncertainty into epistemic and aleatoric components, limiting its utility in settings where model uncertainty (epistemic) must be distinguished from irreducible data noise (aleatoric) for regulatory reporting. We suggest that future work pursue ensemble variants of NGB that inject stochasticity across base learner configurations to recover an approximate epistemic decomposition.

We emphasize that the observed behavior of BNN, BLR, and CNN-BNN reflects a specific combination of implementation choices: mean-field variational inference with a fully factorized Gaussian posterior, standard normal priors, limited training epochs, and the class-imbalance structure of this dataset. Mean-field VI is known to underestimate posterior variance and is susceptible to prior

dominance when the likelihood signal is weak relative to the prior [10,26], as may occur when the minority class represents only 1.6% of observations. More expressive posterior approximations (e.g., normalizing flows, Stein variational gradient descent), informative domain-specific priors, or sampling-based inference (Monte Carlo dropout, Laplace approximation) may yield substantially different results. Reference [55] demonstrate that BNNs with correct regularisation and large-scale data can outperform deterministic models on both performance and calibration; References [56] and [57] provide evidence that more expressive variational families and model averaging substantially improve the reliability of Bayesian deep learning in practice. The conclusions about Bayesian model behavior in this paper are specific to the mean-field VI implementations evaluated here and should not be interpreted as a general verdict on Bayesian deep learning.

6.2. Limitations

Several limitations constrain the generalisability of the findings. First, the dataset originates from a single Norwegian financial institution covering one historical time window; performance and uncertainty characteristics may differ for other portfolios, loan types, regulatory environments, or macroeconomic conditions. The feature set is restricted to 124 consecutive daily account balance observations and does not include credit bureau scores, income, employment status, loan-to-value ratios, or macroeconomic indicators that are standard in industrial credit scoring pipelines. This limits the ability to draw conclusions about feature importance in richer real-world settings and means that the dominance of balance-level features (e.g., *min_balance*, *total_balance*) reflects the informational constraints of this dataset rather than a general finding about credit ML. Conclusions about ‘credit risk machine learning’ broadly should be read as specific to the proprietary account-balance-only setting described herein. Second, the absence of credit bureau variables, income, employment status, and loan-to-value ratios limits the informational content of the feature set; the dominance of aleatoric uncertainty across all models suggests that observed prediction uncertainty partly reflects genuine informational limits rather than modeling deficiencies. Third, bootstrap confidence intervals for AUC and Brier score differences between models were not computed; the reported differences should therefore be treated as descriptive rather than subjected to formal inferential testing. Fourth, the Bayesian models were trained with standard mean-field VI and default prior specifications; more expressive approximate posteriors or informative domain priors might mitigate the posterior collapse observed here, though such investigation was beyond the scope of this benchmarking study.

Fifth, the best-performing preprocessing configuration for each model was identified by test AUC, which introduces an optimistic bias into the headline performance figures reported in Table 3. Accordingly, these figures represent exploratory upper-bound estimates of model capability rather than unbiased out-of-sample performance measures. Readers are encouraged to consult the full 180-configuration table (Table D1 and Appendix C1) to assess the spread of results and evaluate the robustness of cross-model rankings to this selection step. Sixth, the attribution of BNN, BLR, and CNN-BNN failures to posterior collapse relies on behavioral evidence (near-uniform predicted probabilities, low uncertainty dispersion, poor calibration) rather than direct analysis of the variational posterior. ELBO trajectories, KL divergence during training, and posterior-vs-prior parameter comparisons were not monitored, which precludes formal verification of the specific failure mode. Variational inference convergence was assessed indirectly via cross-validation AUC across candidate epoch counts (50 vs. 100); substantially longer training with convergence diagnostics may yield

different results. Richer approximate inference techniques, such as hierarchical posteriors, Monte Carlo ensembles, or priors informed by domain knowledge, have been shown to partially mitigate posterior collapse in related settings [57–59]; their exclusion from this study constitutes a limitation.

Sixth, the attribution of BNN, BLR, and CNN-BNN failures to posterior collapse relies on behavioral evidence (near-uniform predicted probabilities, low uncertainty dispersion, poor calibration) rather than direct analysis of the variational posterior. ELBO trajectories, KL divergence during training, and posterior-vs-prior parameter comparisons were not monitored, which precludes formal verification of the specific failure mode. Variational inference convergence was assessed indirectly via cross-validation AUC across candidate epoch counts (50 vs. 100); substantially longer training with convergence diagnostics may yield different results. Richer approximate inference techniques, such as hierarchical posteriors, Monte Carlo ensembles, or priors informed by domain knowledge, have been shown to partially mitigate posterior collapse in related settings [57–59]; their exclusion from this study constitutes a limitation.

Seventh, PR-AUC (average precision) was not computed. Under a 1.6% default rate, PR-AUC is more informative than ROC-AUC for minority-class discrimination [47]. Future benchmarking studies should report PR-AUC as a primary metric alongside AUC and Brier score.

Eighth, the uncertainty estimates presented in Section 5.3 were not subjected to calibration validation (e.g., uncertainty-error correlation, risk-coverage curves, or selective prediction analysis). Future work should assess whether reported epistemic and aleatoric uncertainty is calibrated and predictive of actual error rates before drawing operational conclusions.

7. Conclusions

This paper presents a multidimensional benchmarking study of probabilistic versus deterministic machine learning for credit default prediction under severe class imbalance (default rate = 1.6%). Across 180 configurations of ten models evaluated on four dimensions, predictive performance, preprocessing robustness, uncertainty quantification, and SHAP interpretability, three empirical conclusions emerge.

First, probabilistic models do not outperform deterministic models on conventional classification metrics. RF and XGB achieve the best AUC and calibration among all models, and NGB matches their AUC while providing informative uncertainty estimates. BNN, BLR, and CNN-BNN exhibit posterior collapse under class imbalance, producing near-uniform predicted probabilities consistent with a failure of the variational approximation to update meaningfully from the prior, though the precise mechanism could not be formally identified from output behavior alone. Performance differences are reported descriptively; formal statistical testing of pairwise AUC comparisons was not conducted and should be interpreted accordingly.

Second, the quality of uncertainty estimates is architecture- and inference-dependent, not a generic property of Bayesian methods. NGB is the only model producing heterogeneous, input-sensitive uncertainty that meaningfully differentiates high- and low-confidence predictions. The Bayesian neural models produce deceptively sharp but uninformative posteriors, challenging the assumption that variational inference reliably propagates data uncertainty into predictions [6, 13, 14].

Third, the preprocessing strategy is a first-order determinant of both performance and model stability. Feature engineering based on rolling statistics and distributional aggregates improves AUC and Brier scores for tree-based models by two to ten percentage points, while the choice of resampling method

induces variance differences of two to three orders of magnitude between tree-based and Bayesian architectures. These results establish that data preprocessing, model architecture, and uncertainty inference must be co-designed rather than treated as independent pipeline stages.

From a practical standpoint, the findings support NGB as a promising candidate for probabilistic credit scoring when uncertainty quantification is operationally required: It delivers competitive discrimination, robust preprocessing stability, and economically coherent feature attributions without the posterior collapse pathology of Bayesian neural models. However, the universal miscalibration documented across all ten models, including the best-performing RF, underscores the need for post-hoc calibration or calibration-aware training before probabilistic outputs are used to set risk thresholds or capital charges in production systems.

Future work should prioritize chronological train-test splits to assess temporal stability, possibly with a train/validation/test experimental design, bootstrap confidence intervals to establish the statistical significance of observed performance differences, formal calibration correction via isotonic regression or temperature scaling, and ensemble variants of NGB to recover epistemic uncertainty decomposition. PR-AUC and expected calibration error (ECE) could be reported alongside AUC and Brier score; DeLong's test or paired bootstrap confidence intervals could be used for formal pairwise model comparisons; and ELBO and KL divergence might be monitored during variational inference training. More expressive variational families or hierarchical priors may also reduce the susceptibility of Bayesian neural models to posterior collapse under severe class imbalance.

Formal statistical comparison of AUC values between models would require DeLong's test [60], paired bootstrap confidence intervals [61], or a Bayesian model comparison framework. Given the identical test set used across all models, DeLong's test for correlated AUC estimates is the most appropriate frequentist procedure; it is recommended for future work. In particular, the apparently similar AUCs of RF, XGB, and NGB (all ≈ 0.80) should not be interpreted as evidence of equivalent performance without formal testing.

The uncertainty estimates presented in Section 4.3 were not subjected to calibration validation (e.g., uncertainty-error correlation, risk-coverage curves, or expected calibration error for probabilistic outputs). Future work should assess whether reported epistemic and aleatoric uncertainty is calibrated and predictive of actual error rates before drawing operational conclusions about uncertainty quality.

Use of AI tools declaration

During the preparation of this work the authors used ChatGPT and Writefull in order to improve language and readability. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Acknowledgments

The authors thank the editor (Sun-Yong Choi) and two anonymous reviewers whose constructive and insightful comments helped us substantially improve the paper.

Conflict of interest

The authors declare there is no conflict of interest.

Author contributions

Engan: Methodology, software, formal analysis, investigation, writing orig. draft.

Hjelkrem: Conceptualization, methodology, validation, investigation, data curation, writing orig. draft, supervision.

Risstad: Conceptualization, methodology, validation, formal analysis, investigation, writing orig. draft, writing rev. and edit., supervision, project administration.

References

1. European Union, *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 on laying down harmonised rules on artificial intelligence and amending certain Union legislative acts (Artificial Intelligence Act)*, 2024. Available from: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>.
2. B. Supervision, Basel committee on banking supervision, *Principles for Sound Liquidity Risk Management and Supervision (September 2008)* (2008). Available from: <https://www.bis.org/publ/bcbs144.htm>.
3. N. Q. K. Le, Y. Y. Ou, Incorporating efficient radial basis function networks and significant amino acid pairs for predicting GTP binding sites in transport proteins, *BMC Bioinf.*, **17** (2016), 501. <https://doi.org/10.1186/s12859-016-1369-y>
4. N. Q. K. Le, T. T. D. Nguyen, Y. Y. Ou, Identifying the molecular functions of electron transport proteins using radial basis function networks and biochemical properties, *J. Mol. Graphics Modell.*, **73** (2017), 166–178. <https://doi.org/10.1016/j.jmgm.2017.01.003>
5. O. N. Nymo, M. Risstad, U. T. Sjøli, T. B. Sæther, Revisiting the spanning hypothesis with machine learning methods: A systematic literature review, *J. Int. Money Financ.*, **167** (2026), 103632. <https://doi.org/10.1016/j.jimonfin.2026.103632>
6. S. Eggen, T. J. Espe, K. Grude, M. Risstad, R. Sandberg, Financial time series uncertainty: A review of probabilistic AI applications, *J. Econ. Surv.*, (2025). <https://doi.org/10.1111/joes.70018>
7. E. S. Gunnarsson, H. R. Isern, A. Kaloudis, M. Risstad, B. Vigdel, S. Westgaard, Prediction of realized volatility and implied volatility indices using AI and machine learning: A review, *Int. Rev. Financ. Anal.*, **93** (2024), 103221. <https://doi.org/10.1016/j.irfa.2024.103221>
8. S. Lessmann, B. Baesens, H. V. Seow, L. C. Thomas, Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research, *Eur. J. Oper. Res.*, **247** (2015), 124–136. <https://doi.org/10.1016/j.ejor.2015.05.030>
9. A. Engan, L. O. Hjelkrem, M. Risstad, A review of credit risk assessment in banking—from deterministic models to probabilistic AI, *Finance Res. Open*, **2** (2026), 100125. <https://doi.org/10.1016/j.finr.2026.100125>

10. Z. Ghahramani, Probabilistic machine learning and artificial intelligence, *Nature*, **521** (2015), 452–459. <https://doi.org/10.1038/nature14541>
11. Y. Gal, Z. Ghahramani, Dropout as a Bayesian approximation: Representing model uncertainty in deep learning, in *International Conference on Machine Learning*, PMLR, (2016), 1050–1059. Available from: <https://proceedings.mlr.press/v48/gal16.html>.
12. Y. Wang, Y. Zhang, Y. Lu, X. Yu, A comparative assessment of credit risk model based on machine learning—A case study of bank loan data, *Procedia Comput. Sci.*, **174** (2020), 141–149. <https://doi.org/10.1016/j.procs.2020.06.069>
13. Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, et al., Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift, *Adv. Neural Inf. Process. Syst.*, **32** (2019). Available from: https://proceedings.neurips.cc/paper_files/paper/2019/file/8558cb408c1d76621371888657d2eb1d-Paper.pdf.
14. M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, et al., A review of uncertainty quantification in deep learning: Techniques, applications and challenges, *Inf. Fusion*, **76** (2021), 243–297. <https://doi.org/10.1016/j.inffus.2021.05.008>
15. D. J. Hand, W. E. Henley, Statistical classification methods in consumer credit scoring: A review, *J. R. Stat. Soc. Ser. A*, **160** (1997), 523–541. Available from: <https://www.jstor.org/stable/2983268>
16. J. N. Crook, D. B. Edelman, L. C. Thomas, Recent developments in consumer credit risk assessment, *Eur. J. Oper. Res.*, **183** (2007), 1447–1465. <https://doi.org/10.1016/j.ejor.2006.09.100>
17. I. R. Berrada, F. Barramou, O. B. Alami, Towards a machine learning-based model for corporate loan default prediction, *Int. J. Adv. Comput. Sci. Appl.*, **15** (2024). <https://doi.org/10.14569/IJACSA.2024.0150357>
18. H. Kvamme, N. Sellereite, K. Aas, S. Sjursen, Predicting mortgage default using convolutional neural networks, *Expert Syst. Appl.*, **102** (2018), 207–217. <https://doi.org/10.1016/j.eswa.2018.02.029>
19. L. O. Hjelkrem, P. E. De Lange, E. Nettet, An end-to-end deep learning approach to credit scoring using CNN+ XGBoost on transaction data, *J. Risk Model Valid.*, **16** (2022), 1–25. <https://doi.org/10.21314/JRMV.2022.018>
20. M. C. Aniceto, F. Barboza, H. Kimura, Machine learning predictivity applied to consumer creditworthiness, *Future Bus. J.*, **6** (2020), 37. <https://doi.org/10.1186/s43093-020-00041-w>
21. L. Coenen, W. Verbeke, T. Guns, Machine learning methods for short-term probability of default: A comparison of classification, regression and ranking methods, *J. Oper. Res. Soc.*, **73** (2022), 191–206. <https://doi.org/10.1080/01605682.2020.1865847>
22. P. E. De Lange, B. Melsom, C. B. Vennerød, S. Westgaard, Explainable AI for credit assessment in banks, *J. Risk Financ. Manag.*, **15** (2022), 556. <https://doi.org/10.3390/jrfm15120556>
23. X. Zhu, Q. Chu, X. Song, P. Hu, L. Peng, Explainable prediction of loan default based on machine learning models, *Data Sci. Manag.*, **6** (2023), 123–133. <https://doi.org/10.1016/j.dsm.2023.04.003>
24. N. G. B. Abrahamsen, E. Nylén-Forthun, M. Møller, P. E. de Lange, M. Risstad, Financial distress prediction in the nordics: Early warnings from machine learning models, *J. Risk Financ. Manag.*, **17** (2024), 432. <https://www.mdpi.com/1911-8074/17/10/432>

25. A. Kendall, Y. Gal, What uncertainties do we need in Bayesian deep learning for computer vision, *Adv. Neural Inf. Process. Syst.*, **30** (2017), 5580–5590. Available from: https://proceedings.neurips.cc/paper_files/paper/2017/file/2650d6089a6d640c5e85b2b88265dc2b-Paper.pdf.
26. C. Blundell, J. Cornebise, K. Kavukcuoglu, D. Wierstra, Weight uncertainty in neural network, in *International Conference on Machine Learning*, PMLR, (2015), 1613–1622. Available from: <http://proceedings.mlr.press/v37/blundell15.pdf>.
27. R. Mbuva, I. Boulkaibet, T. Marwala, Automatic relevance determination Bayesian neural networks for credit card default modelling, preprint, arXiv:1906.06382, 2019. <https://doi.org/10.48550/arXiv.1906.06382>
28. C. Louizos, M. Welling, Multiplicative normalizing flows for variational Bayesian neural networks, in *International Conference on Machine Learning*, PMLR, (2017), 2218–2227. Available from: <http://proceedings.mlr.press/v70/louizos17a/louizos17a.pdf>.
29. A. Ashukha, A. Lyzhov, D. Molchanov, D. Vetrov, Pitfalls of in-domain uncertainty estimation and ensembling in deep learning, preprint, arXiv:2002.06470, 2021. <https://doi.org/10.48550/arXiv.2002.06470>
30. J. Černevičienė, A. Kabašinskas, Explainable artificial intelligence (XAI) in finance: A systematic literature review, *Artif. Intell. Rev.*, **57** (2024), 216. <https://doi.org/10.1007/s10462-024-10854-8>
31. C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, *Nat. Mach. Intell.*, **1** (2019), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
32. M. Nagl, M. Nagl, D. Rösch, Quantifying uncertainty of machine learning methods for loss given default, *Front. Appl. Math. Stat.*, **8** (2022), 1076083. <https://doi.org/10.3389/fams.2022.1076083>
33. A. Niculescu-Mizil, R. Caruana, Predicting good probabilities with supervised learning, in *Proceedings of the 22nd International Conference on Machine Learning*, PMLR, (2005), 625–632. <https://doi.org/10.1145/1102351.1102430>
34. C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, in *International Conference on Machine Learning*, PMLR, (2017), 1321–1330. Available from: <http://proceedings.mlr.press/v70/guo17a/guo17a.pdf>.
35. S. Lundberg, S. Lee, A unified approach to interpreting model predictions, in *Advances in Neural Information Processing Systems 30*, (2017), NIPS.
36. Y. Chen, R. Calabrese, B. Martin-Barragan, Interpretable machine learning for imbalanced credit scoring datasets, *Eur. J. Oper. Res.*, **312** (2024), 357–372. <https://doi.org/10.1016/j.ejor.2023.06.036>
37. T. Hastie, R. Tibshirani, J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2009. Available from: <https://link.springer.com/book/10.1007/978-0-387-84858-7>.
38. N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, *J. Artif. Intell. Res.*, **16** (2002), 321–357. <https://doi.org/10.48550/arXiv.1106.1813>
39. H. He, E. A. Garcia, Learning from imbalanced data, *IEEE Trans. Knowl. Data Eng.*, **21** (2009), 1263–1284. <https://doi.org/10.1109/TKDE.2008.239>

40. K. P. Murphy, *Probabilistic Machine Learning: An Introduction*, MIT Press, 2022. Available from: <https://mitpress.mit.edu/9780262046824/probabilistic-machine-learning/>.
41. L. Breiman, Random forests, *Mach. Learn.*, **45** (2001), 5–32. <https://doi.org/10.1023/A:1010933404324>
42. T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, (2016), 785–794. <https://doi.org/10.1145/2939672.2939785>
43. I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, MIT Press, 2016. Available from: <https://mitpress.mit.edu/9780262035613/deep-learning/>.
44. K. P. Murphy, *Probabilistic Machine Learning: Advanced Topics*, MIT Press, 2023. Available from: <https://mitpress.mit.edu/9780262048439/probabilistic-machine-learning/>.
45. T. Duan, A. Anand, D. Y. Ding, K. K. Thai, S. Basu, A. Ng, et al., NGBoost: Natural gradient boosting for probabilistic prediction, in *International Conference on Machine Learning*, PMLR, (2020), 2690–2700. Available from: <http://proceedings.mlr.press/v119/duan20a/duan20a.pdf>.
46. G. W. Brier, Verification of forecasts expressed in terms of probability, *Mon. Weather Rev.*, **78** (1950), 1–3. Available from: https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml.
47. T. Saito, M. Rehmsmeier, The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets, *PLoS One*, **10** (2015), 1–21. <https://doi.org/10.1371/journal.pone.0118432>
48. B. Baesens, R. Setiono, C. Mues, J. Vanthienen, Using neural network rule extraction and decision tables for credit-risk evaluation, *Manage. Sci.*, **49** (2003), 312–329. Available from: <https://www.jstor.org/stable/4133928>.
49. R. Shwartz-Ziv, A. Armon, Tabular data: Deep learning is not all you need, *Inf. Fusion*, **81** (2022), 84–90. <https://doi.org/10.1016/j.inffus.2021.11.011>
50. J. Van Amersfoort, L. Smith, Y. W. Teh, Y. Gal, Uncertainty estimation using a single deep deterministic neural network, preprint, arXiv:2003.02037, 2020. <https://doi.org/10.48550/arXiv.2003.02037>
51. A. Alemi, B. Poole, I. Fischer, J. Dillon, R. A. Saurous, K. Murphy, Fixing a broken ELBO, preprint, arXiv:1711.00464, 2018. <https://doi.org/10.48550/arXiv.1711.00464>
52. G. Incorvaia, D. Hond, H. Asgari, Uncertainty quantification for machine learning output assurance using anomaly-based dataset dissimilarity measures, in *2023 IEEE International Conference On Artificial Intelligence Testing (AITest)*, IEEE, (2023), 26–33. <https://doi.org/10.1109/AITest58265.2023.00013>
53. E. I. Altman, Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, *J. Finance*, **23** (1968), 589–609. <https://doi.org/10.2307/2978933>
54. L. Thomas, J. Crook, D. Edelman, *Credit Scoring and Its Applications*, SIAM, 2017. Available from: <https://epubs.siam.org/doi/book/10.1137/1.9781611974560>.

55. M. Dusenberry, G. Jerfel, Y. Wen, Y. Ma, J. Snoek, K. Heller, et al., Efficient and scalable Bayesian neural nets with rank-1 factors, preprint, arXiv:2005.07186, 2020. <https://doi.org/10.48550/arXiv.2005.07186>
56. P. Izmailov, S. Vikram, M. D. Hoffman, A. G. G. Wilson, What are Bayesian neural network posteriors really like, in *International Conference on Machine Learning*, PMLR, (2021), 4629–4640. Available from: <http://proceedings.mlr.press/v139/izmailov21a/izmailov21a.pdf>.
57. A. G. Wilson, P. Izmailov, Bayesian deep learning and a probabilistic perspective of generalization, preprint, arXiv:2002.08791, 2022. <https://doi.org/10.48550/arXiv.2002.08791>
58. B. Dai, D. Wipf, Diagnosing and enhancing VAE models, preprint, arXiv:1903.05789, 2019. <https://doi.org/10.48550/arXiv.1903.05789>
59. J. He, D. Spokoyny, G. Neubig, T. Berg-Kirkpatrick, Lagging inference networks and posterior collapse in variational autoencoders, preprint, arXiv:1901.05534, 2019. <https://doi.org/10.48550/arXiv.1901.05534>
60. E. R. DeLong, D. M. DeLong, D. L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach, *Biometrics*, **44** (1988), 837–845. <https://doi.org/10.2307/2531595>
61. R. J. Tibshirani, B. Efron, An introduction to the bootstrap, *Monogr. Stat. Appl. Probab.*, **57** (1993), 1–436. <https://doi.org/10.1201/9780429246593>

Appendix A

A.1. Descriptive statistics

Table A1 presents descriptive statistics for a representative subset of the raw daily account balance features prior to any preprocessing. The full 124-feature descriptive statistics table is available from the authors on request.

Table A1. Descriptive statistics for a subset of the 124 raw daily account balance features for the full dataset before train/test split ($n = 50,819$). Statistics are computed on the pre-standardization values; features are standardized prior to modeling (hence the non-unit standard deviations and wide range reflect the raw scale).

Feature	Count	Mean	Std	Min	25%	50%	75%	Max
dg_1	50,819	−0.0031	1.3139	−13.742	−0.0227	−0.0192	0.0159	261.41
dg_2	50,819	−0.0033	1.3141	−13.742	−0.0226	−0.0191	0.0144	261.41
dg_3	50,819	−0.0017	1.3154	−13.742	−0.0226	−0.0185	0.0175	261.41
dg_10	50,819	0.0004	1.3148	−13.742	−0.0226	−0.0178	0.0261	261.41
dg_30	50,819	0.0008	1.3151	−13.742	−0.0226	−0.0175	0.0265	261.41
dg_60	50,819	0.0002	1.3152	−13.742	−0.0226	−0.0176	0.0249	261.41
dg_90	50,819	−0.0003	1.3148	−13.742	−0.0227	−0.0182	0.0227	261.41
dg_114	50,819	−0.0012	1.3139	−13.742	−0.0227	−0.0194	0.0190	261.41

Appendix B

B.1. Engineered features

Table B1 lists all engineered features by module, with code names matching those used in SHAP plots throughout the paper.

Table B1. Overview of all engineered features grouped by feature engineering module. Code names correspond to feature labels used in SHAP visualizations.

	Code Name	Description
<i>Account statistics</i>		
Total balance	total_balance	Sum of daily balances over the 124-day observation window.
Minimum balance	min_balance	Minimum daily balance over the 124-day period.
Maximum balance	max_balance	Maximum daily balance over the 124-day period.
Mean balance	mean_balance	Mean of daily balances across the 124-day window.
Std. deviation of balance	std_balance	Standard deviation of daily balances across the 124-day window.
<i>Rolling aggregates</i>		
Rolling mean	rolling_mean	7-day rolling mean of daily balances.
Rolling median	rolling_median	7-day rolling median of daily balances.
Rolling std. deviation	rolling_std	7-day rolling standard deviation of daily balances.
Rolling IQR	rolling_IQR	7-day rolling interquartile range of daily balances.
Rolling trend slope	rolling_trend_slope	Slope of the 7-day rolling means.
Rolling trend change count	rolling_trend_change_count	Number of sign changes in the 7-day rolling means.
<i>Variability metrics</i>		
Raw trend change count	raw_trend_change_count	Number of sign changes in daily balance differences.
Proportion increasing	proportion_increasing	Fraction of days with a positive balance change.
Spike count	spike_count	Days with absolute change exceeding twice the standard deviation.
Maximum spike	max_spike	Largest absolute daily change among detected spikes.
Average rolling std. deviation	rolling_std_avg	Mean of the 7-day rolling standard deviations.

Appendix C

C.1. Hyperparameter tuning details

Table C1 reports the hyperparameter search ranges and final values selected by SMBO for all ten models.

Table C1. Hyperparameter tuning overview. Search ranges and final values obtained via sequential model-based optimization (SMBO) with cross-validation AUC as the objective, applied to the training partition ($n = 38,114$) using the Base configuration (no feature engineering or resampling). NB is absent as GaussianNB has no tunable hyperparameters under the default scikit-learn implementation.

	Values searched	Final value
<i>LR</i>		
Penalty	l2, None	None
Regularisation parameter (C)	10^{-4} to 10^2	27.517
Solver	liblinear, saga, lbfgs	lbfgs
Maximum iterations	100, 200, ..., 1000	800
<i>RF</i>		
Number of estimators	50, 100, 150, ..., 500	450
Maximum tree depth	5, 10, 15, ..., 30	10
<i>XGB</i>		
Number of estimators	50, 100, 150, ..., 500	250
Learning rate	10^{-4} to 0.3	0.036
Maximum tree depth	2, 3, 4, ..., 10	2
<i>DNN</i>		
Dense layers	(64, 32), (128, 64)	(64, 32)
Number of epochs	50, 100	50
Batch size	32, 64, 128	128
Dropout rate	0, 0.1, ..., 0.5	0.2
Learning rate	10^{-5} to 10^{-2}	0.000205
Loss function	binary_crossentropy, focal	binary_crossentropy
<i>CNN-DNN</i>		
Convolutional layers	(32,), (64,)	(64,)
Kernel size	3, 5	5
Dense layers	(32,), (64,)	(32,)
Number of epochs	50, 100	50
Batch size	32, 64, 128	128
Dropout rate	0, 0.1, ..., 0.5	0
Learning rate	10^{-5} to 10^{-2}	0.000033
Loss function	binary_crossentropy, focal	binary_crossentropy
<i>BNN</i>		
Bayesian layers	(32, 32), (64, 32)	(32, 32)
Number of epochs	50, 100	50
Batch size	32, 64, 128	64
Learning rate	10^{-5} to 10^{-2}	0.001056
<i>BLR</i>		
Number of epochs	50, 100	50
Batch size	32, 64, 128	128
Learning rate	10^{-5} to 10^{-2}	0.002550
Loss function	binary_crossentropy, focal	binary_crossentropy

(Continued on next page)

Table C1. (Continued.)

	Values searched	Final value
<i>NGB</i>		
Number of estimators	50, 100, 150, ..., 500	400
Learning rate	10^{-4} to 0.3	0.012
<i>CNN-BNN</i>		
Convolutional layers	(32,), (64,)	(32,)
Kernel size	3, 5	5
Bayesian layers	(32,), (64,)	(32,)
Number of epochs	50, 100	50
Batch size	32, 64, 128	128
Dropout rate	0, 0.1, ..., 0.5	0.3
Learning rate	10^{-5} to 10^{-2}	0.000762

Appendix D

D.1. Complete performance results

Table D1 presents training and test set performance for all 180 model configurations. Rows in bold indicate the best configuration per model selected by test AUC; ties are broken by Brier score then accuracy.

Table D1. Training and test set performance for selected model configurations. Bold rows indicate the best configuration per model by test AUC; ties are broken first by Brier score then by accuracy. For configuration details, see Table 2 in the main text.

	Accuracy		Precision		Recall		F1		AUC		Brier
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Test
<i>LR</i>											
Base	0.33	0.32	0.02	0.02	0.88	0.84	0.04	0.04	0.62	0.58	0.24
FE.1	0.50	0.46	0.02	0.02	0.77	0.82	0.05	0.05	0.69	0.69	0.23
FE.2	0.30	0.39	0.02	0.02	0.86	0.85	0.04	0.04	0.67	0.66	0.24
FE.3	0.43	0.43	0.02	0.02	0.89	0.88	0.04	0.04	0.68	0.68	0.24
FE.4	0.47	0.46	0.02	0.02	0.85	0.88	0.04	0.05	0.69	0.69	0.23
FE.5	0.43	0.42	0.02	0.02	0.90	0.88	0.04	0.04	0.68	0.68	0.24
FE.1+SMOTE	0.61	0.57	0.62	0.02	0.77	0.64	0.69	0.05	0.69	0.68	0.23
FE.2+SMOTE	0.55	0.53	0.58	0.02	0.82	0.72	0.68	0.05	0.67	0.68	0.24
FE.3+SMOTE	0.64	0.60	0.61	0.02	0.80	0.61	0.68	0.05	0.70	0.69	0.23
FE.4+SMOTE	0.60	0.57	0.60	0.02	0.79	0.66	0.68	0.05	0.69	0.69	0.23
FE.5+SMOTE	0.62	0.59	0.60	0.02	0.81	0.65	0.68	0.05	0.70	0.69	0.23
FE.1+ADASYN	0.57	0.55	0.59	0.02	0.81	0.72	0.68	0.05	0.69	0.69	0.23
FE.2+ADASYN	0.52	0.51	0.57	0.02	0.85	0.78	0.68	0.05	0.67	0.68	0.24
FE.3+ADASYN	0.61	0.58	0.60	0.02	0.81	0.67	0.69	0.05	0.70	0.69	0.23
FE.4+ADASYN	0.58	0.55	0.59	0.02	0.80	0.69	0.68	0.05	0.69	0.68	0.23
FE.5+ADASYN	0.60	0.57	0.60	0.02	0.81	0.67	0.68	0.05	0.70	0.69	0.23
FE.1+RO	0.48	0.46	0.02	0.02	0.75	0.79	0.05	0.05	0.69	0.68	0.23
FE.2+RO	0.36	0.35	0.02	0.02	0.89	0.89	0.04	0.04	0.67	0.66	0.24
FE.3+RO	0.50	0.49	0.02	0.02	0.85	0.83	0.04	0.04	0.69	0.69	0.23
FE.4+RO	0.48	0.46	0.02	0.02	0.83	0.86	0.04	0.04	0.69	0.68	0.23
FE.5+RO	0.45	0.43	0.02	0.02	0.89	0.88	0.04	0.05	0.68	0.68	0.23
FE.1+RU	0.57	0.54	0.02	0.02	0.65	0.74	0.04	0.05	0.67	0.67	0.23
FE.2+RU	0.53	0.52	0.02	0.02	0.70	0.78	0.04	0.05	0.66	0.66	0.23
FE.3+RU	0.54	0.53	0.02	0.02	0.73	0.74	0.05	0.05	0.68	0.67	0.23
FE.4+RU	0.57	0.53	0.02	0.02	0.66	0.74	0.04	0.05	0.68	0.67	0.23
FE.5+RU	0.55	0.53	0.02	0.02	0.71	0.75	0.04	0.05	0.68	0.68	0.23
<i>RF</i>											
Base	0.79	0.76	0.07	0.04	0.96	0.63	0.14	0.08	0.93	0.79	0.12
FE.1	0.81	0.78	0.08	0.05	0.96	0.64	0.15	0.08	0.94	0.80	0.11
FE.2	0.80	0.77	0.08	0.05	0.96	0.63	0.14	0.08	0.94	0.80	0.12
FE.3	0.80	0.77	0.07	0.04	0.96	0.63	0.13	0.07	0.94	0.79	0.12
FE.4	0.80	0.77	0.07	0.04	0.97	0.64	0.14	0.08	0.94	0.80	0.12
FE.5	0.80	0.77	0.08	0.04	0.97	0.65	0.14	0.08	0.94	0.80	0.12
FE.1+SMOTE	0.77	0.74	0.06	0.04	0.91	0.64	0.12	0.07	0.93	0.79	0.13
FE.2+SMOTE	0.77	0.74	0.05	0.04	0.90	0.64	0.10	0.07	0.93	0.79	0.13

(Continued on next page)

Table D1. (Continued.)

	Accuracy		Precision		Recall		F1		AUC		Brier
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Test
FE.3+SMOTE	0.77	0.73	0.05	0.04	0.90	0.63	0.10	0.07	0.93	0.79	0.13
FE.4+SMOTE	0.77	0.74	0.05	0.04	0.92	0.64	0.10	0.07	0.93	0.79	0.13
FE.5+SMOTE	0.77	0.74	0.06	0.04	0.91	0.64	0.11	0.07	0.93	0.79	0.13
FE.1+ADASYN	0.78	0.74	0.05	0.04	0.91	0.64	0.10	0.07	0.93	0.79	0.13
FE.2+ADASYN	0.77	0.74	0.05	0.04	0.91	0.65	0.10	0.07	0.93	0.79	0.13
FE.3+ADASYN	0.77	0.74	0.05	0.04	0.91	0.65	0.10	0.07	0.93	0.79	0.13
FE.4+ADASYN	0.78	0.74	0.06	0.04	0.90	0.63	0.11	0.07	0.93	0.79	0.13
FE.5+ADASYN	0.77	0.74	0.05	0.04	0.91	0.64	0.10	0.07	0.93	0.79	0.13
FE.1+RO	0.80	0.77	0.07	0.04	0.96	0.65	0.13	0.08	0.94	0.80	0.12
FE.2+RO	0.80	0.77	0.07	0.04	0.96	0.63	0.13	0.07	0.94	0.79	0.12
FE.3+RO	0.80	0.77	0.07	0.04	0.96	0.64	0.13	0.07	0.94	0.79	0.12
FE.4+RO	0.80	0.77	0.07	0.04	0.96	0.65	0.13	0.07	0.94	0.80	0.12
FE.5+RO	0.80	0.77	0.07	0.04	0.96	0.64	0.13	0.07	0.94	0.79	0.12
FE.1+RU	0.77	0.75	0.06	0.04	0.90	0.63	0.12	0.07	0.93	0.79	0.12
FE.2+RU	0.77	0.74	0.06	0.04	0.90	0.64	0.11	0.07	0.93	0.79	0.13
FE.3+RU	0.76	0.74	0.05	0.04	0.91	0.64	0.10	0.07	0.93	0.79	0.12
FE.4+RU	0.77	0.75	0.06	0.04	0.90	0.63	0.11	0.07	0.93	0.79	0.12
FE.5+RU	0.77	0.74	0.06	0.04	0.90	0.63	0.11	0.07	0.93	0.79	0.12
<i>XGB</i>											
Base	0.44	0.44	0.02	0.02	0.90	0.91	0.04	0.04	0.73	0.73	0.20
FE.1	0.63	0.64	0.04	0.04	0.87	0.81	0.07	0.07	0.84	0.80	0.18
FE.2	0.60	0.60	0.04	0.03	0.88	0.82	0.07	0.06	0.82	0.79	0.18
FE.3	0.60	0.60	0.03	0.03	0.87	0.83	0.06	0.06	0.81	0.79	0.19
FE.4	0.60	0.60	0.04	0.03	0.87	0.82	0.07	0.06	0.82	0.80	0.18
FE.5	0.59	0.60	0.03	0.03	0.88	0.83	0.06	0.06	0.81	0.79	0.19
FE.1+SMOTE	0.62	0.62	0.04	0.04	0.87	0.80	0.07	0.07	0.84	0.80	0.18
FE.2+SMOTE	0.58	0.59	0.03	0.03	0.88	0.83	0.06	0.06	0.82	0.79	0.18
FE.3+SMOTE	0.59	0.59	0.03	0.03	0.88	0.83	0.06	0.06	0.81	0.79	0.19
FE.4+SMOTE	0.60	0.60	0.04	0.03	0.87	0.82	0.07	0.06	0.83	0.80	0.18
FE.5+SMOTE	0.59	0.59	0.03	0.03	0.88	0.83	0.06	0.06	0.81	0.79	0.19
FE.1+ADASYN	0.62	0.63	0.04	0.04	0.88	0.81	0.07	0.07	0.84	0.80	0.18
FE.2+ADASYN	0.57	0.57	0.03	0.03	0.88	0.84	0.06	0.06	0.82	0.79	0.18
FE.3+ADASYN	0.58	0.59	0.03	0.03	0.89	0.84	0.05	0.06	0.81	0.79	0.19
FE.4+ADASYN	0.60	0.60	0.04	0.04	0.87	0.81	0.07	0.07	0.83	0.80	0.18
FE.5+ADASYN	0.58	0.59	0.03	0.03	0.88	0.84	0.05	0.06	0.81	0.79	0.19
FE.1+RO	0.63	0.64	0.04	0.04	0.87	0.81	0.07	0.07	0.84	0.80	0.18
FE.2+RO	0.59	0.59	0.03	0.03	0.88	0.83	0.06	0.06	0.82	0.79	0.18
FE.3+RO	0.60	0.60	0.03	0.03	0.87	0.83	0.06	0.06	0.81	0.79	0.18
FE.4+RO	0.60	0.60	0.04	0.03	0.87	0.81	0.07	0.06	0.83	0.80	0.18
FE.5+RO	0.59	0.59	0.03	0.03	0.88	0.83	0.05	0.06	0.81	0.79	0.18
FE.1+RU	0.62	0.62	0.04	0.04	0.87	0.81	0.07	0.07	0.84	0.80	0.18
FE.2+RU	0.59	0.58	0.03	0.03	0.87	0.83	0.06	0.06	0.82	0.79	0.18
FE.3+RU	0.59	0.59	0.03	0.03	0.87	0.83	0.06	0.06	0.81	0.79	0.18
FE.4+RU	0.60	0.60	0.04	0.03	0.88	0.82	0.07	0.06	0.82	0.80	0.18
FE.5+RU	0.59	0.59	0.03	0.03	0.88	0.83	0.06	0.06	0.81	0.79	0.18
<i>NGB (selected configurations)</i>											
Base	0.76	0.64	0.72	0.04	0.87	0.77	0.79	0.07	0.86	0.79	0.18
FE.3	0.77	0.64	0.73	0.04	0.88	0.76	0.80	0.07	0.87	0.79	0.18
FE.3+RO	0.80	0.68	0.76	0.04	0.89	0.76	0.82	0.07	0.89	0.80	0.17
FE.3+SMOTE	0.70	0.62	0.66	0.03	0.90	0.79	0.76	0.06	0.87	0.79	0.18
FE.3+ADASYN	0.71	0.63	0.67	0.03	0.90	0.78	0.77	0.06	0.87	0.79	0.18
FE.3+RU	0.74	0.63	0.70	0.03	0.89	0.77	0.79	0.06	0.87	0.79	0.19
<i>BNN (selected configurations)</i>											
Base	0.96	0.96	0.00	0.00	0.00	0.00	0.00	0.00	0.62	0.60	0.04
FE.3	0.50	0.16	0.45	0.02	0.75	0.96	0.56	0.04	0.71	0.69	0.27
FE.3+SMOTE	0.54	0.13	0.49	0.02	0.77	1.00	0.60	0.04	0.71	0.71	0.27
FE.5+ADASYN	0.53	0.12	0.47	0.02	0.80	1.00	0.55	0.04	0.70	0.72	0.26
FE.3+ADASYN	0.51	0.14	0.47	0.02	0.78	0.99	0.58	0.04	0.70	0.70	0.28
FE.3+RU	0.70	0.68	0.04	0.03	0.75	0.66	0.08	0.05	0.77	0.68	0.22

Note: DNN, CNN-DNN, BLR, NB, and CNN-BNN full configuration tables follow the same structure. Full table available from authors.



AIMS Press

©2026 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0>)